利用概念化的少样本短文本分类研究

沈炜域　刘奇飞

(中国人民公安大学信息技术与网络安全学院　北京　100038)

摘　要:[目的/意义]旨在为用户和管理者的短文本分类管理提供参考。[方法/过程]利用开放知识库完成词粒度的概念化,将 CWE 预训练得到的词嵌入与实例的概念化表示拼接合成文本表示,并利用相似度的计算预测待标注短文本的类别。[结果/结论]结果表明了在少样本的情况下,该方法分类效果优于实验涉及的其他文本分类模型。

关键词:少样本学习;词嵌入;概念化;文本分类

中图分类号:TP391.1　　**文献标识码**:A　　**Adoi**:10.3969/j.issn.1005-8095.2018.12.002

Research on Few-shot Short Text Classification via Conceptualization

Shen Weiyu　Liu Qifei

(School of Information Technology and Network Security, People's Public Security University of China, Beijing 100038)

Abstract: [**Purpose/significance**] The paper is to provide reference for users' and managers' short text classification management. [**Method/process**] The paper uses external knowledge base to conceptualize word granularity, concatenates the words pre-trained with CWE and conceptualization of living examples to be a text, and finally predicts the category of the short text with similarity calculation. [**Result/conclusion**] The experimental results demonstrate the proposed method achieves better effectiveness than other models concerned in conditions of few-shot text.

Keywords: few-shot learning; word embedding; conceptualization; text classification

0　引言

短文本广泛存在于互联网的各个角落，包括视频弹幕、直播间评论、游戏聊天、智能语音助手等。每天产生的海量短文本数据在辅助用户画像、提升推荐性能的同时,也对色情、涉政、暴恐等多类内容的审查带来了很大的挑战。虽然目前许多基于关键字的搜索方式能够有不错的查准率，但是由于缺乏语义理解，对于一些存在词语变种的文本往往被错分或漏查。如何有效地对短文本进行分类管理,为用户或管理者提供有用信息变得非常重要。

本文提出一种利用外部知识库在少量标注样本的情况下预测新文本的类别的方法。首先,基于不同词性成分对类别的典型度进行排序，进而自动选择出类内具有强典型性的样本;其次,为了融合更多知识，本文利用外部知识库中丰富的概念性知识获得文本的更多语义信息,同时提升消歧的能力;最后,将新样本文本标注为与之语义相似度最高的样本的类别。本文第 3 部分将详细描述模型所有步骤,并在实验部分比较了本文模型与 2 种基线模型在同一数据集的分类效果。

1　相关研究

在短文本分类这样的监督学习任务中，使用恰当的方法表示非结构化的文本样本是关键的环节。传统的文本处理大部分是根据词频和逆向文档频率将文本表示成一个稀疏的向量。经过多年发展,这种简单高效的模型依然得到了广泛的使用，但这种模型忽略了文本的语义信息。随着基于神经网络的表示学习的发展，基于分布假说的语言模型为许多下游的自然语言处理任务带来了新的可能。许多词嵌入模型如 GloVe[1]、字符增强词嵌入模型(Character-enhanced Word Embedding,CWE)[2]等从大规模语料上训练得到的词向量表示为词语带来了一定的语义信息。而对于句子表示,文献[3]提出一种利用预训练的词嵌入的加权平均表征句子级文本的无监督表示方法,使用词频估计给句中的每个词向量赋予权重,然后使用 PCA 或者 SVD 方法移除句向量中的无关部分,进而得到句子表示。

在得到文本表示之后，如何将这些表示映射到

收稿日期:2018-08-07

作者简介:沈炜域(1993—),男,2016 级硕士研究生,主要研究方向为文本挖掘、内容安全;刘奇飞(1994—),男,2016 级硕士研究生,主要研究方向为信息安全、大数据。

相应的类别中去是文本分类任务的重点问题。在文本分类任务中，经典的机器学习算法具有严格的数学推导过程，依然可以达到很高的分类精度。基于分类器的算法，如支持向量机[4]、逻辑回归[5]等被广泛应用在话题分类和情感分析等任务中。此外还有一些基于相似度计算的方法被用到分类或聚类中，如文献[6]使用改进的词移距离(WMD)算法进行聚类，在新闻评论数据集上取得了很好效果。

上述算法往往需要很多标注样本投入训练或参与相似度的排序，但是在许多垂直领域中，标注样本往往很少。为了解决在少量标注数据的情况下解决分类问题，文献[7]提出了称为“原型网络”的方法，通过计算和每个类别的原型表达的距离来进行分类。文中定义原型向量为类内支持点集合的向量平均，每个类别的原型向量都能够表达类内样本在样本空间中的映射。通过这种方式将分类问题变成在样本空间中的最近邻问题。本文沿用“原型”的叫法，将类内具有代表性的实例和概念称为原型实例和原型概念。

为了让短小的文本加入更多语义信息，文献[8]等证明了依赖额外的知识能够帮助机器“理解”词与词之间的关系。概念化作为一类基于显性语义(Explicit Semantic Analysis，ESA)[9]的短文本理解方法，旨在建立文本中的实例到其概念的映射。显性语义模型最大的特点就是它所产生的短文本向量表示不仅是可用于计算的，也是人类可以理解的，每一个维度都有明确的含义，通常是一个明确的“概念”[10]。文献[11-12]借助知识库，推导出实例的概念分布。文献[13]借助中国知网分析词项间的语义关系，提出一种基于中国知网语义相似度的文本相似度加权算法，并对该算法进行中文文本分类实验。

短文本字数少，且常常不遵循语法规则，没有足够的信息进行统计推断。经过实验统计发现，短文本中的谓词、形容词等成分为短文本的分类提供了重要的依据，因此本文使用拼接的方式保留短文本部分词性成分，同时基于不同词性成分对类别的重要程度进行排序，选择出类内具有强典型性的样本。为了融合更多知识，本文利用中文维基百科和中文Probase[14]对短文本中的实例进行概念化，通过将概念化得到的结果与短文本中的其他成分的向量表示进行拼接，进而计算新样本文本与各个标注样本之间的语义相似度，最终将新样本文本标注为与之相似度最高的样本的类别。

2 利用知识库语义扩展的短文本分类方法

2.1 原型选择

选择类内具有强典型性的样本作为该类的表示是提高分类正确率的重要因素。虽然在少数样本的情况下，人工参与选择同样能够达到较好的效果，但是本文依然使用一种简单但行之有效的方法来选择最佳的样本。文献[15]使用 LDA 主题模型[16]在长文本(新闻)上计算词-类别分布并排序来推导原型表示。由于短文本字词较少，因此本文提出的方法是通过不同词性中的词对于类别的重要程度排序，并选择 TopN 作为各类别的相应词性的原型集。

通过词性标注，从短文本中抽取出谓词与形容词，计算各个类别中每个谓词和形容词与类别的重要程度。需要特别说明的是，单独分析谓词和形容词的原因是在短文本分类任务中，在信息极少的情况下，谓词和形容词对消歧起着十分重要的作用。图 1 展示了 21 种意图识别数据集词性分布。

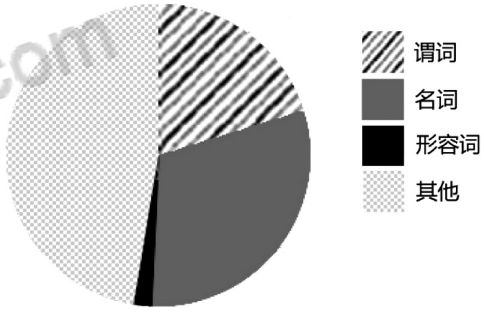

图 1 21 种意图识别数据集词性分布

本文使用 $f_c(v)$ 表征谓词 v 在类别 c 上的重要程度，使用 $f_c(a)$ 表征形容词在类别 c 上的重要程度。$f_c(v)$ 的计算方法如下：

$$f_c(v)=\frac{n_c(v)}{\sum_k n_c(v_k)}\times\left(1+log\frac{|C|+1}{N+1}\right). \quad (1)$$

其中，$n_c(v)$ 为谓词 v 在类别 c 中出现的频次，为类别 c 中所有特征项出现的频次之和，N 为所有类别中含有谓词 v 的类别个数，$|C|$ 为类别总数。形容词 $f_c(a)$ 之于类别的典型度的计算方法与计算 $f_c(v)$ 的方法相同。

对每一个类别中的谓词和形容词进行重要程度计算，分别得到表 1 和表 2 所示的重要程度得分。

对每个类别内的谓词和形容词的重要程度进行排序，选取最大的 n 个谓词和 m 个形容词作为原型谓词集与原型形容词集。利用原型谓词集与原型形容词集，本文从各个类别中选择出如“浏览器”“春熙路”等实例。将这一系列实例作为原型实例，利用中

表 1　类内谓词的重要程度得分

类别 c	谓词 v	$f_c(v)$
视频	播放	0.2301
	放映	0.2079
	…	…
小说	推荐	1.3283
	搜索	0.3310
	…	…
…	…	…

表 2　类内形容词的重要程度得分

类别 c	形容词 a	$f_c(a)$
视频	恐怖	2.7587
	幸福	1.8391
	…	…
小说	灵异	2.6566
	最火	1.3283
	…	…
…	…	…

文维基百科和中文 Probase 获得每个原型实例所属的概念,并将这些概念作为原型概念,计算每一个实例之于每一个概念的典型度。典型度度量了某个实例是因为某个概念被提及的可能性。通过实例对于其概念的典型度的计算,得到维度为实例的概念个数的概念分布。

2.2　概念分布的计算

本文使用概念分布作为实例的语义表示。H_e 是一个维度为概念数 n 的向量,表示实例 e 的概念分布。

$$H_e=<\omega_1,\omega_2,\cdots,\omega_n>. \tag{2}$$

每个维度为实例 e 之于概念 c_i 的典型度:

$$\omega_i=p(c_i\mid e)\quad i\in\{1,2,\cdots,n\}. \tag{3}$$

根据实例 e 与概念 c_i 在 Web 网页中的共现次数计算典型度:

$$p(c_i\mid e)=\frac{d(c_i,e)}{d(e)}\quad i\in\{1,2,\cdots,n\}. \tag{4}$$

其中,$d(c_i,e)$表示概念 c_i 与实例 e 的共现频数,$d(e)$表示实例 e 的出现频数。

2.3　文本表示

本文使用 CWE 训练得到词嵌入表示。在 CWE 模型中,对于上下文中的词语的表征,一部分来自于词粒度的向量表示,还有一部分来自于这些词语中的字粒度的向量表示。具体的上下文词语的嵌入表示 x_i 计算方式如下:

$$x_i=\frac{1}{2}\left(w_i+\frac{1}{N_i}\sum\nolimits_{j=1}^{N_i}z_j\right). \tag{5}$$

其中,w_i 是该词的词粒度嵌入表示,N_i 是该词中的字符的个数,z_j 是该词第 j 个字符的字粒度嵌入表示。

受到张量模型[17]的启发,为了强调不同词性的词在语义合成中的不同作用,本文将不同词性的词表示进行拼接。对于未在样本中出现的词性,使用类内该词性的加权平均值作为该样本对应词性的向量。

文本合成算法的输入为:词嵌入表示$\{v_w:w\in V\}$,词的类内概率估计$\{P_{C_i}(w):w\in V\}$,类别 C_i 的原型谓词集 $P^*_{C_i}$。类别 C_i 原型形容词集 $A^*_{C_i}$。对于每一个文本样本,根据其包含的词性计算相应的向量表示。若样本中不存在谓词,则使用类内谓词的加权平均值作为该样本的谓词部分的表示向量 v_p,即

$$v_p=\frac{1}{|P^*_{C_i}|}\sum_{w\in|P^*_{C_i}|}\frac{P_{C_i}(w)}{\sum P_{C_i}(w)}v_w.$$

若样本中不存在形容词,则对应的形容词向量 v_a 为

$$v_a=\frac{1}{|A^*_{C_i}|}\sum_{w\in|A^*_{C_i}|}\frac{P_{C_i}(w)}{\sum P_{C_i}(w)}v_w.$$

本文利用搜索引擎在 Web 网页中提取出短文本中出现的实例与各概念的共现频率,从而计算出该实例之于各个概念的典型度向量作为概念分布 H_e。将得到的 v_p、v_a 与 H_e 进行直接拼接得到 v_d 作为短文本的最终表示。

2.4　类别预测

预测新文本类别时,计算测试文本向量v_{test} 与原型样本向量(带有原型谓词或原型形容词的样本)表示集合$\{v_d:d\in D\}$中的每一个向量计算相似度算 $sd=similarity(v_d,v_{\text{test}})$,将相似度最大的原型样本对应的类别作为该测试样本的类别。

3　实验

3.1　数据与实验设置

实验使用的数据是由讯飞整理的中文人机交互对话数据集的 1 个子集,数据集包括了 21 种垂直领域的用户意图分类。实验样本文本的平均长度为 12.1 个字符。实验选取类内典型度排序最大的 5 个谓词和 5 个形容词作为原型谓词与原型形容词。对于缺少形容词或谓词的样本,本文使用类内形容词词嵌入加权平均或类内谓词词嵌入加权平均进行填充。

经过原型选择,实验得到的原型实例数量分布如图 2 所示,其中,纵坐标表示原型实例个数,横坐标表示不同类别。

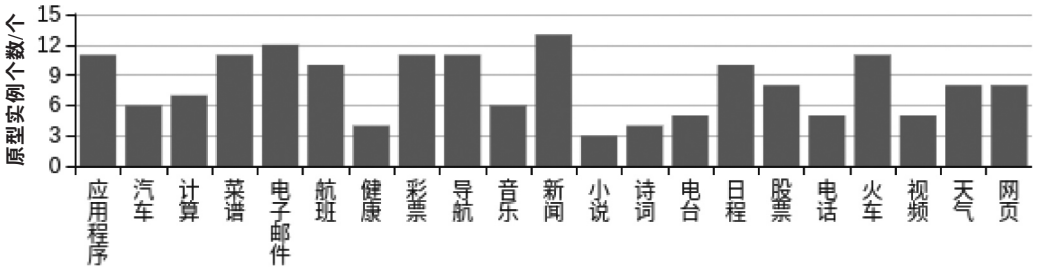

图 2　原型实例分布直方图

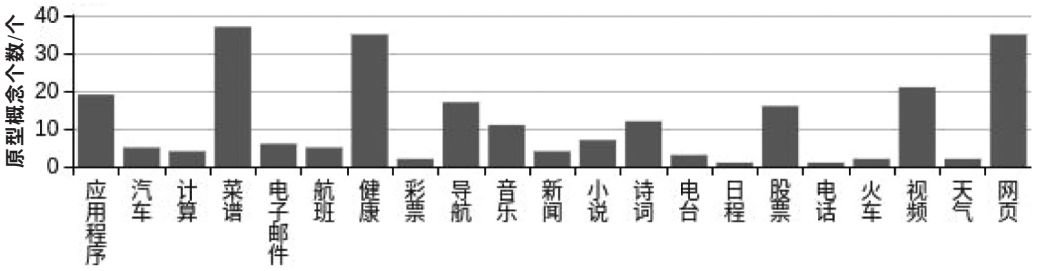

图 3　原型概念分布直方图

通过中文维基百科和中文 Probase，实验得到上述原型实例所涉及的 245 个原型概念，其数量分布如图 3 所示，其中，纵坐标表示原型概念个数，横坐标表示不同类别。

实验将中文维基百科的内容页面数据导入全文检索引擎 ElasticSearch 集群，利用其全文检索能力加快典型度计算的速度。由于实例的多样性，知识库无法保证涵盖所有实例。为了获得未被知识库收录的实例，本文利用检索引擎在维基百科的内容页面中检索该实例，并将返回结果中相关度 TopK（本文实验 K=2）的实例概念作为该实例的概念。实验在预测阶段采用 cosine 距离度量文本表示间的相似度。

本文对比了 2 种文本分类任务中常用的算法，包括利用词移距离计算文本相似度的最近邻算法[18]，以及在此类任务中较常用且性能优异的支持向量机。

为了探究词嵌入维度对准确率的影响，本文在维基百科 dump 数据语料上使用 CWE 分别迭代 30 轮（skip-gram 模式，窗口大小为 10）预训练得到 100 维、200 维、300 维 3 个词嵌入矩阵，将 3 种长度的词嵌入应用于实验中。此外，为了探究概念分布向量维度对准确性的影响，本文尝试直接使用标注类别的名称作为概念，经过典型度的计算得到长度等于类别数的向量作为概念分布。

3.2　实验结果与评价

实验使用宏 F1-score 和正确率（accuracy）作为分类性能评价指标。包括本文算法在内的 3 种算法的实验结果如表 3 所示。

表 3　算法性能指标

算法	F1 值	正确率
SVM（BoW）	0.779	0.784
WMD+KNN（300dim）	0.762	0.763
本文方法（100dim）	0.741	0.746
本文方法（200dim）	0.759	0.759
本文方法（300dim）	0.783	0.784
计算类别分布的方法	0.594	0.597

通过对实验结果的分析，由于中文人机交互对话数据样本较少，其中的实例几乎不重复出现，而且一种意图类别的谓词变化较多，这使得基于 BoW 的 SVM 仅能使用得到特征项较少。同样利用预训练词嵌入的 WMD+KNN 算法对于有歧义的实例如“放一首小苹果”（真实类别“音乐”被预测为“食谱”类）和单实例如“小兵张嘎”（真实类别“视频”被预测为“股票”类）在不加知识库的情况下也被错分。

通过使用不同维度的词嵌入表示，对比实验的结果表明更大维度的词嵌入能给文本表示带入更多语义信息，这有助于提高分类的准确性。由于数据集提供了每个类别准确的真实意图，本文尝试直接使用类别的名称作为概念计算典型度，但最终的分类准确性远低于其他算法。通过对错分样本的个案分析，本文认为其原因是一些样本中提及的实例属于多义词，其作为不同概念的实例时，典型度仅和其所属的概念有关。例如：在真实类别“应用程序”中出现的实例“skype”属于电话类应用程序，是一种网络电话工具，其对于“电话”的典型度高于其对于“应用程

序”的典型度。

4　结论

本文提出了一种预训练词嵌入与知识库相结合的少样本短文本分类方法。本文利用中文维基百科和中文 Probase 获取实例的所属概念,同时在大语料上利用 CWE 预训练词嵌入,将短文本样本中非实例的词向量的加权平均结果与实例的概念分布进行拼接获得最终的文本表示。预测时将测试样本标记为与其余弦距离最小的样本的类别。实验结果表明,本文的方法在文本长度短小的意图分类任务上的准确率高于传统文本分类算法。

对于垂直领域的短文本分类任务,命名实体识别以及实体对齐对分类准确率具有很大的影响。在接下来的研究当中,将考虑将上述 2 种任务与意图分类任务进行联合学习,避免错误向下游传播,从而保证分类的准确性。

参考文献

[1] PENNINGTON J,SOCHER R,MANNING C. Glove:Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg,PA:ACL. 2014:1532-1543.

[2] CHEN X,XU L,LIU Z,et al. Joint Learning of Character and Word Embeddings[C]//International Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2015:1236-1242.

[3] ARORA S,LIANG Y,MA T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings[EB/OL]. [2017-10-17]. https://openreview.net/forum?id=SyK00v5xx.

[4] 张华鑫. 基于 SVM 的文本分类研究[J]. 情报探索,2015(5):133-135.

[5] 李平,戴月明,王艳. 基于混合卡方统计量与逻辑回归的文本情感分析[J]. 计算机工程,2017,43(12):192-196,202.

[6] 官赛萍,靳小龙,徐学可,等. 基于 WMD 距离与近邻传播的新闻评论聚类[J]. 中文信息学报,2017,31(5):203-214.

[7] SNELL J,SWERSKY K,ZEMEL R. Prototypical networks for few-shot learning[C]//Advances in Neural Information Processing Systems 30. New York:Curran Associates,2017:4080-4090.

[8] WANG Z,ZHAO K,WANG H,et al. Query Understanding through Knowledge-Based Conceptualization[C]//International Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2015:3264-3270.

[9] EGOZI O,MARKOVITCH S,GABRILOVICH E. Concept-Based Information Retrieval Using Explicit Semantic Analysis [J]. ACM Transactions on Information Systems,2011,29(2):1-34.

[10] 王仲远,程健鹏,王海勋,等. 短文本理解研究[J]. 计算机研究与发展,2016,53(2):262-269.

[11] SONG Y,WANG H,WANG Z,et al. Short text conceptualization using a probabilistic knowledgebase[C]//Proceedings of the twenty-second international joint conference on artificial intelligence-volume volume three. Palo Alto,CA:AAAI Press,2011:2330-2336.

[12] HUA W,WANG Z,WANG H,et al. Short text understanding through lexical-semantic analysis[C]//International Conference on Data Engineering 2015. Piscataway,NJ:IEEE Press,2015:495-506.

[13] 刘怀亮,杜坤,秦春秀. 基于知网语义相似度的中文文本分类研究[J]. 数据分析与知识发现,2015,31(2):39-45.

[14] LIANG J,XIAO Y,WANG H,et al. Probase+:Inferring Missing Links in Conceptual Taxonomies [J]. IEEE Transactions on Knowledge & Data Engineering,2017,29(6):1281-1295.

[15] BAILEY K,CHOPRA S. Few-Shot Text Classification with Pre-Trained Word Embeddings and a Human in the Loop[EB/OL]. [2018-04-05] https://arxiv.org/abs/1804.02063.

[16] BLEI D M,NG A Y,JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research,2012(3):993-1022.

[17] COECKE B,SADRZADEH M,CLARK S. Mathematical foundations for a compositional distributional model of meaning[EB/OL]. [2010-03-13]. https://arxiv.org/abs/1003.4394.

[18] KUSNER M,SUN Y,KOLKIN N,et al. From word embeddings to document distances[C]//Proceedings of the 32nd International Conference on Machine Learning. Cambridge,MA:MIT Press,2015:957-966.

阅读此文的还阅读了：

1. 英汉被动句在概念化中的对比研究
2. 利用LIS系统样本时间流转功能保证时限性样本检测质量的探讨
3. 概念,而非概念化
4. 土地集约利用的“绍兴样本”
5. 基于LDA特征扩展的短文本分类方法研究
6. 概念化的主体意识性：语言应用的认知研究
7. 英语“good”和汉字“好”的概念化研究
8. 多研究些故事,少谈些IP
9. 姜堰打造湿地保护利用示范样本
10. 工具遗留样本及其利用
11. 少来点“研究研究”
12. 时间概念化的隐喻机制
13. 有辅助信息可利用时的分层抽样下样本轮换研究
14. 词汇新构式创生概念化和再概念化路径研究
15. 材料剩余少就是利用率高吗
16. 土壤资源化利用的美国样本
17. 利用酚醛泡沫塑料制作样本足迹研究
18. 利用胶带粘面捺印样本手印方法的研究
19. 研究结构转型的区域样本
20. 利用photoshop分离指印与文字混合样本的方法研究
21. 概念化与存在句方向的认知研究
22. 利用概念化的少样本短文本分类研究
23. 少依赖朋友,多利用敌人
24. 利用“落后的优势”少走弯路
25. 外语研究中结构方程模型的概念化

26. 基于本体网络概念化单元信息研究

27. 利用Excel函数求解样本协方差矩阵

28. 少依赖朋友,多利用敌人

29. Ontology领域概念化模型的研究

30. 概念化的主观性:认知视点的日语语义研究

31. 土壤资源化利用的美国样本

32. 文本解读的样本研究

33. 利用多旋翼无人飞行器采集覆冰样本的探讨

34. 时间概念化的转喻认知研究

35. 我们的思维也被概念化了

36. 多研究些故事，少谈些IP

37. 中动结构的概念化机制研究

38. 英汉语量度矫正概念化和动态性概念化对比

39. 心智模拟概念化——量词“把”的语义概念化研究

40. “捉-放-捉”--感受利用样本估计总体

41. “捉-放-捉”--感受利用样本估计总体

42. 教育研究中的样本容量

43. 煤炭清洁化利用的淄博样本

44. 大城少村:山东城镇化实践样本

45. “台风”的概念化研究

46. 基于主题模型的短文本分类研究

47. 笔迹检验中案后笔迹样本的收集和利用

48. 概念化、再概念化与中国话语权构建

49. 无表盘的概念化手表

50. 人的城镇化苏州样本研究