计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150742
53(2):262-269,2016

短文本理解研究

王仲远[1,2]　程健鹏[2,3]　王海勋[4]　文继荣[1]
[1](中国人民大学信息学院　北京　100872)
[2](微软亚洲研究院　北京　100080)
[3](牛津大学计算机科学学院　英国牛津　OX1 3QD)
[4](Facebook　美国加利福尼亚州门洛帕克市　94025)
(zhy.wang@microsoft.com)

Short Text Understanding: A Survey

Wang Zhongyuan[1,2], Cheng Jianpeng[2,3], Wang Haixun[4], and Wen Jirong[1]
[1](*School of Information, Renmin University of China, Beijing* 100872)
[2](*Microsoft Research Asia, Beijing* 100080)
[3](*Department of Computer Science, Oxford University, OXford, UK OX1 3QD*)
[4](*Facebook, Menlo Park, CA, USA* 94025)

Abstract　Short text understanding is an important but challenging task relevant for machine intelligence. The task can potentially benefit various online applications, such as search engines, automatic question-answering, online advertising and recommendation systems. In all these applications, the necessary first step is to transform an input text into a machine-interpretable representation, namely to "understand" the short text. To achieve this goal, various approaches have been proposed to leverage external knowledge sources as a complement to the inadequate contextual information accompanying short texts. This survey reviews current progress in short text understanding with a focus on the vector based approaches, which aim to derive the vectorial encoding for a short text. We also explore a few potential research topics in the field of short text understanding.

Key words　knowledge mining; short text understanding; conceptualization; semantic computing

摘　要　短文本理解是一项对于机器智能至关重要但又充满挑战的任务.这项任务有益于众多应用场景,如搜索引擎、自动问答、广告和推荐系统.完成这些应用的首要步骤是将输入文本转化为机器可以诠释的形式,即帮助机器"理解"短文本的含义.基于这一目标,许多方法利用外来知识源来解决短文本中语境信息不足的问题.通过总结短文本理解领域的相关工作,介绍了基于向量的短文本理解框架.同时,探讨了短文本理解领域未来的研究方向.

关键词　知识挖掘;短文本理解;概念化;语义计算

中图法分类号　TP391

短文本理解是一项对于机器智能至关重要的任务.其在知识挖掘领域有很多潜在应用,如网页搜索、在线广告、智能问答等.为了完成这些任务,先前的研究往往使用一些知识库系统,如 Freebase,

收稿日期:2015-08-10;**修回日期**:2015-11-19
基金项目:国家"九七三"基础研究发展计划基金项目(2014CB340403);中央高校基本科研业务费专项资金(14XNLF05)
This work was supported by the National Basic Research Program of China (973 Program) (2014CB340403) and the Fundamental Research Funds for the Central Universities (14XNLF05).

Yago 等为机器“装备”知识.这些知识库大多包含大量实体以及与之相关的事实.基于这些事实,机器可以通过查询的方式获取输入问题的答案.然而,如图 1 所示,在机器回答问题前,首先需要解决的是“理解”问题,这也是这一过程中的最大挑战.

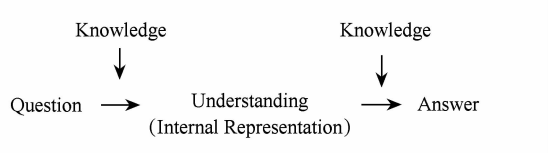


Fig. 1 Question-answering with knowledge requires understanding.

图 1 基于知识的问答过程

对于人类而言,理解问题十分简单.这是由于人类具有“思维”,能够积累知识并做出推断.例如,给出 2 个查询语句“band for wedding”和“wedding band”,人类可以清楚地判断前者指的是一项“婚礼乐队服务”,而后者是“结婚戒指”.

然而,自动化的短文本理解是一项充满挑战的任务.与长文本不同,短文本通常不遵循语法规则,并且长度短、没有足够的信息量来进行统计推断,机器很难在有限的语境中进行准确的推断.此外,由于短文本常常不遵循语法,自然语言处理技术(如词性标注和句法解析等)难以直接应用于短文本分析.

传统的基于短文本的应用常通过枚举和关键词匹配的方式避免“理解”这一任务.例如,在自动问答系统中,一个关于问题和答案匹配的列表可被事先构建,这样在线查询时只需对列表中的条目进行匹配.随着近年来自然语言处理技术的发展,主流的搜索引擎正逐渐从基于关键词的搜索向文本理解过度.例如,给出“apple ipad”这个短文本,机器需要明白“apple”所指为品牌名而不是水果.

许多相关工作[1-3]证明,自动化的短文本理解需要依赖额外的知识.这些知识可以帮助机器充分挖掘短文本中词与词之间的联系,如语义相关性.例如,在英文查询“premiere Lincoln”中,“premiere”是一个重要的信息,表明“Lincoln”在这里指的是“电影”;同样,在“watch harry potter”中,正因为“watch”的出现,“harry potter”的含义可被鉴定为“电影”或“DVD”,而不是“书籍”.但是,这些关于词汇的知识(例如“watch”的对象通常是“电影”)并没有在短文本中明确表示出来,因而需要通过额外的知识源获取.

本文根据所需知识源的属性,将短文本理解模型分为 3 类:隐性(implicit)语义模型、半显性(semi-explicit)语义模型和显性(explicit)语义模型.其中,隐形和半显性模型试图从大量文本数据中挖掘出词与词之间的联系,从而应用于短文本理解.相比之下,显性模型使用人工构建的大规模知识库和词典辅助短文本理解.

从另一个角度而言,短文本理解模型在文本分析上的粒度也有差异.部分方法直接模拟短文本的表示方式,因此本文将其归为“文本”粒度.其余大多方法则以词为基础.这些方法首先推出每个词的表示,然后使用额外的合成方式推出短文本的表示.本文将这些方法归为“词”粒度.图 2 展示了所有短文本理解方法在知识源属性和粒度的二维坐标轴中对应的位置.这些方法将被逐一讨论.

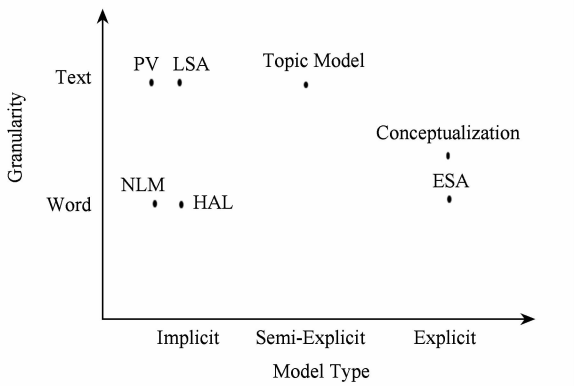


Fig. 2 Models viewed on type-granularity graph.

图 2 不同模型的属性和粒度

1 短文本理解方法

1.1 隐性(implicit)语义模型

隐性语义模型产生的短文本通常表示为映射在一个语义空间上的隐性向量.这个向量的每个维度所代表的含义人们无法解释,只能用于机器计算.以下将介绍 4 种代表性的隐性语义模型.

1) 隐性语义分析模型.最早的基于隐性语义的文本理解框架为隐性语义分析(latent semantic analysis, LSA)[1],也被称为 latent semantic indexing (LSI).LSA 旨在用统计方法分析大量文本,从而推出词与文本的含义表示.其思想核心是在相同语境下出现的词具有较高的语义相关性.具体而言,LSA 构建一个庞大的词与文本的共现矩阵.对于每个词向量,它的每个维度都代表一个文本;对于每个文本向量,其每个维度代表一个词.通常,矩阵每项的输入是经过平滑或转化的共现次数.常用的转化方法为 TF-IDF.最终,LSA 通过奇异值分解(SVD)的

方法将原始矩阵降维. 在短文本的情境下,LSA 有 2 种使用方式:首先,在语料足够多的离线任务上,LSA 可以直接构建一个词与短文本的共现矩阵,从而推出每个短文本的表示;其次,在训练数据较小的情境下,或针对线上任务(针对测试数据),可以事先通过标准的 LSA 方法得到每个词向量,然后使用额外的语义合成方式获取短文本向量.

2) 超空间模拟语言模型. 一个与 LSA 类似的模型是超空间模拟语言模型(hyperspace analogue to language model, HAL)[4]. HAL 与 LSA 的主要区别在于前者是更加纯粹的"词模型". HAL 旨在构建一个词与词的共现矩阵. 对于每个词向量,它的每个维度代表一个"语境词". 模型统计目标词汇与语境词汇的共现次数,并经过相应的平滑或转换(如 TF-IDF, pointwise mutual information 等)得到矩阵中每个输入的值. 通常,语境词的选取有较大的灵活性. 例如,语境词可被选为整个词汇,或者除停止词外的高频词[5]. 类比 LSA,在 HAL 中可以根据原始向量的维度和任务要求选择是否对原始向量进行降维. 由于 HAL 的产出仅仅为词向量,在短文本理解这一任务中需采用额外的合成方式(如向量相加)来推出短文本向量.

3) 神经网络语言模型. 近年来,随着神经网络和特征学习的发展,传统的 HAL 逐渐被神经网络语言模型(neural language model, NLM)[6-9]取代. 与 HAL 通过明确共现统计构建词向量的思想不同,NLM 旨在将词向量当成待学习的模型参数,并通过神经网络在大规模非结构化文本的训练来更新这些参数以得到最优的词语义编码(常被称作 word embedding).

最早的概率性 NLM 由 Bengio 等人提出[6],其模型使用前向神经网络(feedforward neural network)根据语境预测下一个词出现的概率. 通过对训练文本中每个词的极大似然估计,模型参数(包括词向量和神经网络参数)可使用误差反向传播算法(BP)进行更新. 此模型的一个缺点在于仅仅使用了有限的语境. 后来,Mikolov 等人[7]提出使用递归神经网络(recurrent neural network)来代替前向神经网络,从而模拟较长的语境. 此外,原始 NLM 的计算复杂度很高,这主要是由于网络中大量参数和非线性转换所致. 针对这一问题,Mikolov 等人[8]提出 2 种简化(去掉神经网络权重和非线性转换)的 NLM,即 continuous bag of words(CBOW)和 Skip-gram. 前者通过窗口语境预测目标词出现的概率,而后者使用目标词预测窗口中的每个语境词出现的概率.

另一类非概率性的神经网络以 Collobert 和 Weston 的工作[9]为代表. 其模型 Senna 考虑文本中的 n 元组. 对每个 n 元组中某个位置的词(例如中间词),模型选取随机词来代替该词,从而产生若干新的 n 元组作为负样本. 在训练中,通过一个简单的神经网络为 n 元组打分,训练目标为正样本得分 s^+ 与负样本得分 s^- 间的最大间隔排序损失(max-margin ranking loss),如式(1)所示:

$$\sum_{n\text{-gram}} \max(0, 1 - s^+ + s^-). \tag{1}$$

总而言之,NLM 同 HAL 相似,所得到的词向量并不能直接用于短文本理解,而需要额外的合成模型依据词向量得到短文本向量.

4) 段向量. 段向量(paragraph vector, PV)[10]是另一种基于神经网络的隐性短文本理解模型. PV 可被视作文献[8]中 CBOW 和 Skip-gram 的延伸,可直接应用于短文本向量的学习. PV 的核心思想是将短文本向量当作"语境",用于辅助推理(例如根据当前词预测语境词). 在极大似然的估计过程中,文本向量亦被作为模型参数更新. PV 的产出是词向量和文本向量. 对于(线上任务中的)测试短文本,PV 需要使用额外的推理获取其向量. 图 3 比较了 CBOW、Skip-gram 和 2 种 PV 的异同.

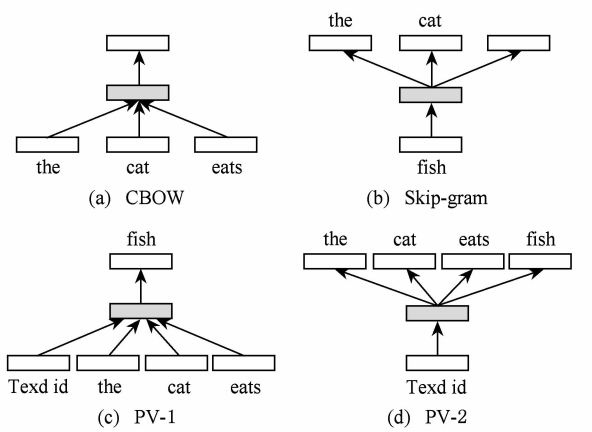


Fig. 3 CBOW, Skip-gram and two variations of PV.

图 3 CBOW, Skip-gram 和 2 种 PV

1.2 半显性(semi-explicit)语义模型

半显性语义模型产生的短文本表示方法,也是一种映射在语义空间里的向量. 与隐性语义模型不同的是,半显性语义模型的向量的每一个维度是一个"主题(topic)". 这个主题通常是一组词的聚类. 人们可以通过这个主题猜测这个维度所代表的含义;但是这个维度的语义仍然不是明确的、可解释的. 半显性语义模型的代表性工作是主题模型(topic models).

LSA 尝试通过线性代数(奇异值分解)的处理方式发现文本中的隐藏语义结构,从而得到词和文本的特征表示;而主题模型则尝试从概率生成模型(generative model)的角度分析文本语义结构,模拟"主题"这一隐藏参数,从而解释词与文本的共现关系.

最早的主题模型 PLSA(probabilistic LSA)为 LSA 的延伸,由 Hofmann 提出[11]. PLSA 假设文本具有主题分布,而文本中的词从主题对应的词分布中抽取. 以 d 表示文本,w 表示词,z 表示主题(隐藏参数),文本和词的联系概率 $p(d,w)$ 的生成过程可被表示为:

$$p(d,w) = p(d)\sum_{z\in Z} p(w \mid z)p(z \mid d). \quad (2)$$

虽然 PLSA 可以模拟每个文本的主题分布,然而其没有假设主题的先验分布(每个训练文本的主题分布相对独立),它的参数随训练文本的个数呈线性增长,且无法应用于测试文本.

一个更加完善的主题模型为 LDA(latent Dirichlet allocation)[12]. LDA 从贝叶斯的角度为 2 个多项式分布添加了狄利克雷先验分布,从而解决了 PLSA 中存在的问题. 在 LDA 中,每个文本的主题分布为多项式分布 $Mult(\theta)$,其中 θ 从狄利克雷先验 $Dir(\alpha)$ 抽取. 同理,对于主题的词分布 $Mult(\varphi)$,其参数 φ 从狄利克雷先验 $Dir(\beta)$ 获取. 图 4 对比了 PLSA 和 LDA 的盘子表示法(plate notation).

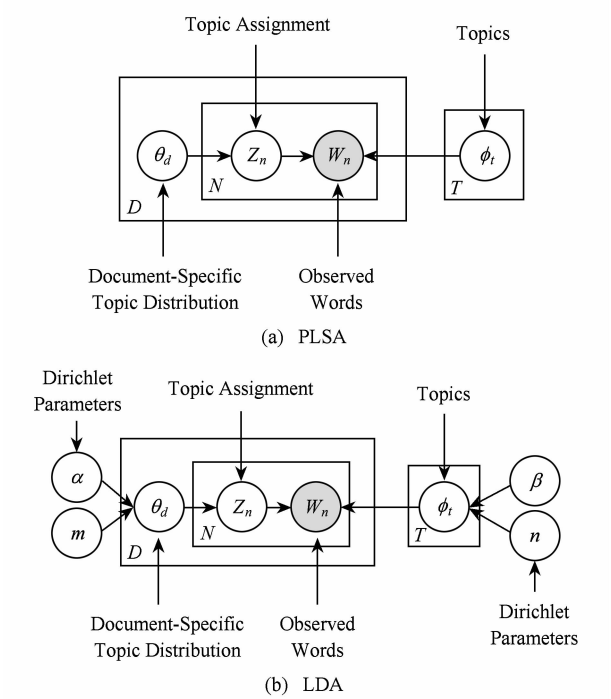


Fig. 4 Plate notations of PLSA and LDA in comparison.

图 4 PLSA 和 LDA 盘子表示法比较

总之,通过采用主题模型对短文本进行训练,最终可以获取每个短文本的主题分布,以作为其表示方式. 这种表示方法将短文本转为了机器可以用于计算的向量.

1.3 显性(explicit)语义模型

近年来,随着大规模知识库系统的出现(如 Wikipedia,Freebase,Probase 等),越来越多的研究关注于如何将短文本转化成人和机器都可以理解的表示方法. 这类模型称之为显性语义模型. 与前 2 类模型相比,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机器计算的,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(concept)". 这意味着机器将短文本转为显性语义向量后,人们很容易就可以判断这个向量的质量,发现其中的问题,从而方便进一步的模型调整与优化.

1) 显性语义分析模型. 在基于隐性语义的模型中,向量的每个维度并没有明确的含义标注. 与之相对的是显性语义模型,向量空间的构建由知识库辅助完成. 显性语义分析模型(explicit semantic analysis,ESA)[13]同 LSA 的构建思路一致,旨在构建一个庞大的词与文本的共现矩阵. 在这个矩阵中,每个输入为词与文本的 TF-IDF. 然而,在 ESA 中词向量的每个维度代表一个明确的知识库文本,例如 Wikipedia 文章(或标题). 此外,原始的 ESA 模型没有对共现矩阵进行降维处理,因而产生的词向量具有较高维度. 在短文本理解这一任务中,需使用额外的语义

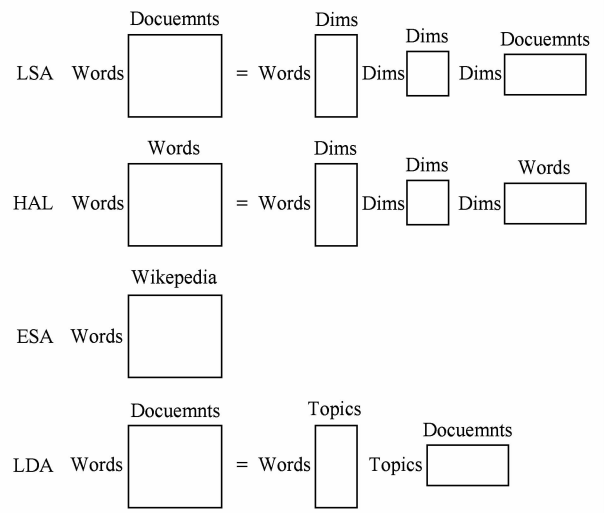


Fig. 5 LSA, HAL, ESA, LDA in comparison.

图 5 LSA, HAL, ESA, LDA 比较

合成方法推导出短文本向量.图5比较了LSA,HAL,ESA和LDA在本质上的区别与联系.

2)概念化.另一类基于显性语义的短文本理解方法为概念化(conceptualization)[2-3,14-15].概念化旨在借助知识库推出短文本中每个词的概念分布,即将词按语境映射到一个以概念为维度的向量上.在这一任务中,每个词的候选概念可从知识库中明确获取.例如,通过知识库Probase[16],机器可获悉apple这个词有"水果"和"公司"这2个概念.当apple出现在"apple ipad"这个短文本中,通过概念化可分析得出apple有较高的概率属于"公司"这个概念.

最早的概念化方法由Song等人提出[2].其模型使用知识库Probase,获取短文本中每个词与概念间的条件概率 $p(concept|word)$ 和 $p(word|concept)$,从而通过朴素贝叶斯的方法推出每个短文本的概念分布.这一单纯基于概率的模型无法处理由语义相关但概念不同的词组成的短文本(如"apple ipad").为解决无法识别语境的问题,Kim等人[14]对Song的模型做出了改进.新的模型使用LDA主题模型,分析整条短文本的主题分布,进而计算 $p(concept|word,topic)$.

另一个基于Probase的短文本理解框架为Hua等人提出的LexSA(lexical semantic analysis)[15].LexSA将短文本理解系统化为分词、词性标注和概念识别3个步骤,并在每个步骤使用新的模型消除歧义.在分词和词性标注环节,作者分别使用图模型推出短文本的最优分词方式和词的词性;在概念识别环节,每个词被表示成以概念为维度的向量.为了进一强调LexSA中各环节的相互作用关系,Wang等人[3]提出为短文本构建统一的候选词关系图,并使用随机漫步(random walk)的方法推导出最优的分词、词性和词的概念.

2 模型粒度分析

本节将深入讨论第1节的短文本理解模型在文本分析粒度上的差异,并从应用层面论证不同方法的适用性.

2.1 文本粒度模型

首先,文本粒度的模型包含LSA,LDA和PV.这些模型均尝试直接推导出短文本的向量表示作为模型的输出.在LSA中,通过构建一个词与文本的共现矩阵,每个文本可用以词为维度的向量表示.类似地,LDA试图模拟文本的生成过程.作为结果,可得到每个文本的主题分布.PV通过神经网络推测(inference)的方式获取文本向量的最优参数.上述模型所得的文本向量均可以直接用于与这些文本相关的任务,如文本分类[17-18]、聚类[19]、摘要生成[20].值得注意的是,LSA同时输出词向量.因而在短文本数量不足的情况下,可以先采用基于大量完整文本的LSA获取词向量,再通过额外的合成方法获取短文本向量.对于LDA和PV而言,其模型亦可以通过额外的文本训练,然后应用于短文本.

2.2 词粒度模型

同LSA,LDA和PV相比,其他模型(LSA,NLM,ESA等)均属于词粒度的模型.这是由于这些模型的产出仅为词向量.针对短文本理解这一任务,必须使用额外的合成手段来推出短文本的表示.例如,在文献[21-25]工作中,作者均利用词向量推导出文本表示,并用于后续的文本相似度判断、文本复述、情感分析等任务.这里的一个特例为概念化模型.由于概念化可以直接基于语境推出短文本中每个词的概念,这样的输出方式已经可以满足机器短文本理解的需求.因而概念化虽属于词粒度的模型但并不需要额外的文本合成.

2.3 文本合成

如何通过词向量获取任意长度的文本向量(包括短文本)是时下流行的一个研究领域.根据复杂度的不同,文本合成方法可被大致分为代数向量模型[5,21-23,25]、张量模型[26-28]和神经网络模型[7,24,29-32].

1)代数运算模型.最早的合成模型由Mitchell和Lapata[21]提出.其模型使用逐点的(point-wise)向量相加的方式从词向量推出文本向量.虽然这一基于"词袋"的方法忽略了句子中的词序("cat eats fish"和"fish eats cat"将有相同的表示),事实表明其在很多自然语言处理任务上有着不错的效果,且其常常被用作复杂模型的基准[23].类似的代数运算模型还有逐点的向量乘积[5,21-22]以及乘法与加法的结合运算[24].

2)张量模型.张量模型[26-27]为代数运算模型的延伸.其试图强调不同词性的词在语义合成中的不同角色.例如在"red car"这个词组中,形容词"red"对名词"car"起修饰作用;而在"eat apple"中,动词"eat"的角色好比作用于"apple"的函数.从这个角度而言,将不同词性的词均表示为同等维度的向量过于简化.因而,在张量模型中,不同词性的词被表示为不同维度的张量,整个句子的表示方式以张量乘法

的形式获取.目前,张量模型的最大挑战是如何获取向量与张量的映射关系[28].

3) 神经网络模型.时下最为流行的文本合成模型为基于神经网络的模型,如 recursive neural network (RecNN)[29-30],recurrent neural network(RNN)[5],convolutional neural network (CNN)[31-32] 等.在这些模型中,最基本的合成单元为神经网络.通常的形式为神经网络根据输入向量 $\boldsymbol{x}_1$,$\boldsymbol{x}_2$ 推出其组合向量 $\boldsymbol{y}$:

$$\boldsymbol{y}=f(\boldsymbol{W}[\boldsymbol{x}_1:\boldsymbol{x}_2]+b), \tag{3}$$

其中,$\boldsymbol{W}$ 和 b 为神经网络参数,$[\boldsymbol{x}_1:\boldsymbol{x}_2]$为 2 个输入向量相连,$f$ 为非线性转换.

在具体的文本合成中,不同的神经网络模型的构造不同.例如,RecNN 依赖于语法树开展逐层的语义合成,它无法被用于短文本.相比之下,RNN(序列合成)和 CNN(卷积合成)都可以快速通过词向量推导出短文本向量.

3 未来的研究展望

短文本理解是对机器智能至关重要的一项任务.针对机器智能的特质,自动化的短文本理解可定义为:将文本转化为任何机器可以获取其含义并进行进一步计算的编码形式.基于此,大量先前工作(如 LSA,NLM,LDA 等)通过挖掘文本数据中的隐藏信息,获取词与词、词与文本之间的联系,从而获取短文本编码.与此同时,另一方向的研究(ESA、概念化)使用知识库来获取明确的词汇语义知识,从而辅助短文本理解.尽管,何为最有效的短文本解释方式仍有待探索,本文将尝试从 2 个方面讨论短文本理解领域的未来工作.

3.1 语义知识网

知识对短文本的理解不可或缺.传统的知识库(如 WordNet,Freebase,Yago 等)往往包含大量与实体相关的事实,但机器无法直接根据这些非黑即白的事实进行线上推测.针对这个问题,未来的一个趋势是探索概率性的语义知识网在短文本理解上的应用.

语义知识网旨在帮助机器"理解"人类的交流方式,而不仅仅是记录事实片段.例如,NELL,Probase 等新兴的知识库均属于语义网.这些网络以自然语言为导向,且通常包含大量的统计信息,如词与词的共现关系.下文将以 Probase 为例简述语义网在短文本理解任务上的作用.

Probase 基于 16.8 亿网页构建,它包含了大量基于 Hearst 模式获取的 isA 关系,例如"Obama" isA "president","China" isA "developing country"等.与传统的知识库不同,Probase 的语义网记载了实体与概念之间的概率.因此,在基于 Probase 的短文本理解工作中,机器可以通过语义网中的概率进行在线推导.

目前,虽然已经有一些基于语义知识网来进行短文理解的工作[2-3,14-15],但这些工作仍然比较初步,多是基于一些观察所构建的模型,缺乏系统性理论支持.未来工作可深入探索语义网在解读短文本工作上的应用,构建一套完备的理论模型.

3.2 显性知识和隐性知识的结合

从另一个角度而言,如上文所述,机器可获取的知识包含了显性知识和隐性知识.未来工作应着重探索二者的结合以完善短文本含义的表示方式.

1) 显性知识改进隐性模型

显性的知识库可以用来完善隐性的空间向量.换言之,向量应以某种方式反应知识库中实体间的关系.例如,Bian 等人的工作[33]使用 WordNet 中的词汇关系作为限制来辅助 NLM 的训练,使得这些词汇关系(如同义词关系)能够在训练所得的词向量中得以体现.

2) 隐性知识改进显性模型

隐性的空间向量可以帮助提高概念化的准确性.例如,Cheng 等人[34]使用改进的 NLM 将 Probase 的实体和概念以及文本中的其他词均映射至统一的向量空间.在这样的设置下,对于某一实体词,其语境词与概念的相关性可以很容易地使用空间距离度量.这一结合语境判断概念的方法有潜力提升概念化的效果.

如图 6 所示,未来工作应围绕强调显性和隐性知识的联系,构建能够更准确体现真实的词与概念语义的向量空间,提升图 1 中理解(即通过短文本推导出机器内部表示)这一环节的准确性.

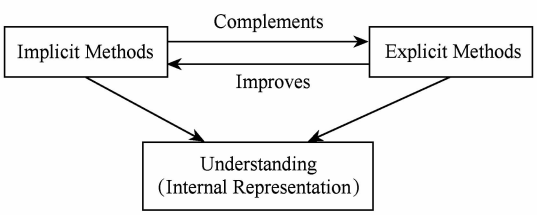


Fig. 6 Combining explicit and implicit knowledge in short text understanding in the future.

图 6 未来结合显性和隐性知识辅助短文本理解

4 结束语

综上所述,随着短文本数据迅猛增长,短文本理解研究是近年来一个研究热点,这也是基于关键字的搜索技术达到一定瓶颈之后的必然选择.本文从隐性模型、半显性模型以及显性模型的角度,介绍了目前比较流行的短文本理解语义模型,并深入阐述了它们之间的关联与不同.基于这些分析,本文尝试提出了未来在短文本理解上的2种研究方向,供相关研究人员参考.

参 考 文 献

[1] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis [J]. Journal of the Association of Information Sience, 1990, 41(6): 391-407

[2] Song Y, Wang H, Wang Z, et al. Short text conceptualization using a probabilistic knowledgebase [C] // Proc of the 22nd Int Joint Conf on Artificial Intelligence (IJCAI). Palo Alto, CA: AAAI, 2011: 2330-2336

[3] Wang Z, Zhao K, Wang H, et al. Query understanding through knowledge-based conceptualization [C] //Proc of the 24th Int Joint Conf on Artificial Intelligence (IJCAI). Palo Alto, CA: AAAI, 2015: 3264-3270

[4] Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence [J]. Behavior Research Methods, Instruments, & Computers, 1996, 28(2): 203-208

[5] Turney P D, Pantel P. From frequency to meaning: Vector space models of semantics [J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188

[6] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. The Journal of Machine Learning Research, 2003, 3(2): 1137-1155

[7] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C] //Proc of the 11th Annual Conf of the Int Speech Communication Association. New York: ACM, 2010: 1045-1048

[8] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Computing Research Repository, 2013 [2015-12-30]. http://arxiv.org/pdf/1301.3781.pdf

[9] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C] //Proc of the 25th Int Conf on Machine Learning (ICML). New York: ACM, 2008: 160-167

[10] Le Q V, Mikolov T. Distributed representations of sentences and documents [C] //Proc of the 31st Int Conf on Machine Learning(ICML). Palo Alto, CA: AAAI, 2014: 1188-1196

[11] Hofmann T. Probabilistic latent semantic indexing [C] // Proc of the 22nd Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1999: 50-57

[12] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3(1): 993-1022

[13] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis [C] //Proc of the 20th Int Joint Conf on Artificial Intelligence (IJCAI). San Francisco, CA: Morgan Kaufmann, 2007: 1606-1611

[14] Kim D, Wang H, Oh A. Context-dependent conceptualization [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence (IJCAI). Palo Alto, CA: AAAI, 2013: 2654-2661

[15] Hua W, Wang Z, Wang H, et al. Short text understanding through lexical-semantic analysis [C] //Proc of the 31st Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2015: 495-506

[16] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding [C] //Proc of the 2012 ACM Int Conf on Management of Data (SIGMOD). New York: ACM, 2012: 481-492

[17] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys (CSUR), 2002, 34(1): 1-47

[18] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620

[19] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization [C] //Proc of the 26th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2003: 267-273

[20] Geiss J. Latent semantic sentence clustering for multi-document summarization [D]. Cambridge, UK: University of Cambridge, 2011

[21] Mitchell J, Lapata M. Vector-based models of semantic composition [C] //Proc of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2008: 236-244

[22] Erk K, Padó S. A structured vector space model for word meaning in context [C] //Proc of the 2008 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2008: 897-906

[23] Blacoe W, Lapata M. A comparison of vector-based representations for semantic composition [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: ACL, 2012: 546-556

[24] Hermann K M, Blunsom P. The role of syntax in vector space models of compositional semantics [C] //Proc of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2013: 894-904

[25] Fyshe A, Talukdar P, Murphy B, et al. Documents and dependencies: An exploration of vector space models for semantic composition [C] //Proc of the 17th Conf on Computational Natural Language Learning. Stroudsburg, PA: ACL, 2013: 84-93

[26] Coecke B, Sadrzadeh M, Clark S. Mathematical foundations for distributed compositional model of meaning [J]. Linguistic Analysis, 2010, 36: 345-384

[27] Baroni M, Zamparelli R. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. Stroudsburg, PA: ACL, 2010: 1183-1193

[28] Kartsaklis D. Compositional operators in distributional semantics [J]. Springer Science Reviews, 2014, 2(1/2): 161-177

[29] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks [C] // Proc of the 28th Int Conf on Machine Learning (ICML). Madison, WI: Omnipress, 2011: 129-136

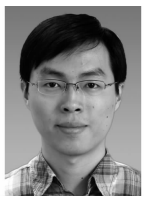
[30] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1631-1642

[31] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences [C] // Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014: 655-665

[32] Kim Y. Convolutional neural networks for sentence classification [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751

[33] Bian J, Gao B, Liu T Y. Knowledge-powered deep learning for word embedding [M] //Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2014: 132-148

[34] Cheng J, Wang Z, Wen J, et al. Contextual text understanding in distributional semantic space [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2015: 133-142



Wang Zhongyuan, born in 1985. PhD. Researcher at the Microsoft Research Asia. His research interests include short text understanding, knowledgebase, NLP, and machine learning.

Cheng Jianpeng, born in 1990. PhD candidate of Oxford University. His main research interests include machine learning and natural language understanding.

Wang Haixun, born in 1972. Research scientist at Facebook. Before he joined Facebook, he was research scientist at Google Research and senior researcher at Microsoft Research Asia. His main research interests include text analytics, NLP, knowledgebase, and graph data management.

Wen Jirong, born in 1972. Professor at the Renmin University of China. His main research interests include big data management & analytics, information retrieval, data mining and machine learning.