

文章编号: 1003-0077(2019)04-0037-11

基于领域知识的增强约束词向量

王恒升^{1,2}, 刘 通¹, 任 晋¹

(1. 中南大学 机电工程学院, 湖南 长沙 410083;
2. 中南大学 高性能复杂制造国家重点实验室, 湖南 长沙 410083)

摘 要: 词向量是一种词语的数字化的表达。基于神经网络模型, 利用语料中词语之间的上下文关系这一约束条件, 通过大量训练得到词向量。词向量在表达词的语义上的表现给人以无限的希望与想象空间, 基于词向量的文本分类、人机对话、智能检索等得到了广泛的研究。该文针对校园信息查询的特定应用, 建立了所涉及词语的分类本体, 除了利用语料中词语上下文关系外, 还将本体知识作为约束条件进行词向量的训练, 增强了词向量的语义表达。基于 skip-gram 模型, 采用多任务的神经网络训练方法, 在自己收集的语料上训练得到了针对领域的词向量。实验表明, 基于领域知识的增强约束词向量能够更准确地表达词的语义信息。

关键词: 增强约束词向量; 语义表达; 本体知识

中图分类号: TP391 **文献标识码:** A

Constraint-enhanced Word Embedding Based on Domain Knowledge

WANG Hengsheng^{1,2}, LIU Tong¹, REN Jin¹

(1. College of Mechanical & Electrical Engineering, Central South University, Changsha, Hunan 410083, China;
2. State Key Laboratory for High Performance Complex Manufacturing, Central South University, Changsha, Hunan 410083, China)

Abstract: For the design of a specific application of natural language based dialog system, i. e. campus information inquiry system, this paper proposes a method of improving word embedding for the expressiveness of semantic meanings. In addition to employing the word contexts in the training of word embedding, the domain specific knowledge is also introduced into the model training to enhance the expressiveness of word embedding. The knowledge about the application is organized into an ontology which was incorporated into word embedding through multi-task training of neural network model adapted from skip-gram, which is both a kind of constraint and a kind of enhancement to the word embedding. Experiments show the validness of the proposed embedding.

Keywords: constraint-enhanced word embedding; semantic expression; ontology knowledge

0 引言

自然语言是人类生产生活中长期积累形成的, 用于表达情感、意图的工具以及记录、传播知识的载体。文字是语言的基本构成, 是记录语言的符号体系^[1]。用信息技术的术语来讲, 自然语言是一种典型的信息系统。

计算机及信息技术的发展, 为自然语言的计算机处理奠定了基础。按照认识论的分类, 计算机自

然语言处理可分为基于理性主义的方法和基于经验主义的方法。研究工作的早期阶段以依靠人类知识构建各种语言及语法规则的理性主义研究为主, 形成基于规则的句法分析和语义分析技术^[2]。随着处理自然语言的规模不断扩大, 自然语言作为伴随人类进化过程而不断进化的一种信息系统, 其复杂性特征和各种语言现象层出不穷, 基于理性主义的方法很快遇到了瓶颈, 导致无法处理大规模的真实文本^[3]。基于统计学理论的经验主义方法, 利用大规模语料库, 使用概率统计的方法建立语言模型^[4], 取

得了意想不到的成功。由于基于经验主义的方法在理论的完备性上存在不足,从长远来讲,要解决这个问题,必须将两种方法结合起来,彼此取长补短,才能相得益彰^[5]。本文基于这一思想,针对自然语言处理的一个特殊应用场合(限定场合的对话系统),将理性知识融入词向量(一种统计语言模型)中,对词向量建模过程进行干预,得到本文称为“增强约束词向量”的文本模型。经实验测定,增强约束词向量具有更强的词语表达能力,针对本文的应用,能更准确地得到自然语言的语义信息。

本文后续安排如下:第1节介绍应用背景,第2节阐述增强约束词向量;第3节介绍基于本体知识的增强约束词向量;第4节实验验证并分析实验结果;第5节总结全文,得出结论。

1 应用背景

计算机的应用进军到自然语言领域可以说是科学家雄心勃勃努力的结果,这一工作与人工智能有密切的联系,以至于图灵测试把人机自然语言对话作为通过人工智能检验的一个标准。基于自然语言的人机对话系统按照用途可分为开放型和领域任务型两种。开放型人机对话系统不针对具体问题,是一种开放式的对话,常用作聊天机器人;领域任务型人机对话系统往往针对某一场景,旨在帮助人类解决某一方面的实际应用问题。

本文将自然语言处理任务应用于大学校园的信息查询,针对某大学的一个集教学、科研和实验于一体的综合大楼,提供信息查询、路径导航等功能,是一种领域任务型人机对话系统的应用。(其中的信息查询系统本身不是本文的主要内容,不做详细介绍)

语义理解是人机对话系统的关键部分,其任务是对人的自然语言输入指令进行意图识别及要素提取。大多数学者采用改进语义理解模型的方法提高语义理解的准确性,如 Xu^[6] 等将 TriCRF 模型与卷积神经网络结合,用于航班预定对话的语义理解,相比标准 TriCRF 模型在所用数据集上取得了更好的效果。Zhang^[7] 等基于循环神经网络提出一种联合模型,同时进行对话的意图识别与要素抽取,在所用对话数据集中,两个任务均取得了最佳的效果。尽管利用改进的模型取得了不错的效果,然而大多数模型的特征输入来源于传统方法:手工标注、one-hot 或基于词频的表示(如 TF-IDF 等)。手工标注费时费力,one-hot 与基于词频的表示虽能自动构

建,但无法考虑特征间的语义相关性,这个缺点成了进一步提高模型性能的瓶颈。

基于 Harris 假设^[8]“具有相似上下文的词语,其语义是相似的”,Bengio^[9] 等于 2003 年提出经典的词向量训练模型——神经网络语言模型,将词的表示向量化,同时保证了语义接近的词语其词向量也是接近的。自此,众多研究学者开始利用带有语义信息的词向量作为文本的特征表达,以克服传统特征表示方法的缺点,提高任务效果。冯艳红^[10] 等基于词向量技术得到文本特征向量,采用 CRF 方法实现了领域术语识别,相比于传统的 TF-IDF 特征,提高了领域术语识别的精度;Liao^[11] 等利用含有文本主题信息的词向量作为输入特征,在中文情感分析任务中取得了良好的效果。

词的向量表达为自然语言处理打开了一个通道,吸引了大批的研究者。但由于自然语言本身的复杂性,已有的词向量训练模型得到的词向量往往表达力有限,词向量的实际应用效果还有待提高,寻找更好的词向量表达成为一个关键问题,将人类关于自然语言方面的知识显式地融入词向量中成为许多学者努力的方向。一类方式是融入通用语言学方面的知识,如字词的形态学特性(如汉语中的偏旁部首、英语中的词根、前后缀等)、句子的语法规则(如词性、英语中的比较级、单复数、时态等)、词语的语义特性(如文本数据库 WordNet, Freebase, Probase 等提供的词的关联关系)^[12],或者词的情感特性^[13];文献[14]利用句法知识,将动词、名词信息加入词向量的学习过程中,得到更准确的表达。另一类方式是将领域知识显式地融入词向量表达中,如 Liu^[15] 等利用本体构建学术论文的领域知识,在训练过程中融入学术论文的语义关系,最大化上下文约束与领域知识约束。Chen^[16] 等利用 UMLS(统一医学语言系统)作为额外的知识库,结合大量与医疗相关的未标注文章,生成医学领域词向量。Taghipour^[17] 等提出一种改进的词向量模型(adapted word embeddings),在词向量学习过程中增加更具有区别性的领域信息,得到针对于特定领域(金融、体育)的词向量,提高了词语消歧系统的准确性。

本文的研究属于上述第二类,针对校园信息查询对话系统的特殊应用,建立该系统基于本体的知识库,改进 skip-gram 的训练模型,将该应用的本体知识融入词向量的训练过程中,改造词向量的基分布,在词向量的表达中体现该应用的知识,提高该对话系统对用户提问的理解的准确性,实现更为自然

流畅的对话过程。

2 增强约束词向量

训练词向量的基本思想可以理解为以语料库中词语之间的上下文关系为约束条件,对神经网络模型中的参数进行优化,其中的一部分模型参数就构成了词语的数字化表达,将其表示成向量形式,就是词语的向量表达。这种模型看似简单,但经过海量的模型训练,得到惊人的表达效果:相近含义的词语会在向量空间中相对集中,具有相似关系的词语之间的向量差也会得到相近的向量。下面先简单介绍一下词向量的训练方法,然后介绍本文提出的基于领域本体知识的增强约束的词向量训练方法。

2.1 skip-gram 方法

Word2Vec 是 Google 公司 2013 年开放的、目前应用广泛的训练词向量的软件工具,是基于 Mikolov 等^[18]所提出神经网络训练模型的实现,包括 CBOW 和 skip-gram 两种模型。本文采用 skip-gram 模型。

Skip-gram 模型由输入层、投影层和输出层组成,如图 1 所示。skip-gram 模型由前馈神经网络语言模型(feedforward neural net language model, NNLM)改进而来。与 NNLM 不同的是,skip-gram 去掉了非线性隐藏层,投影层由全部词语共享。

整个模型是一种全连接神经网络,这里将输入层与投影层之间的权值矩阵 W 称为词向量矩阵, $W \in \mathbb{R}^{k \times V}$; 投影层与输出层之间的权值矩阵称为辅助矩阵 W' , $W' \in \mathbb{R}^{V \times k}$, k 表示词向量维数, V 表示词典大小。

模型输入是中心词 w_t 的 one-hot 表示 w_t ($w_t \in \mathbb{R}^{V \times 1}$), 它的第 t 个元素为 1, 其余为 0。

投影层对输入进行式(1)所示的操作,取出词向量矩阵 W 中对应中心词 w_t 的第 t 列的列向量 v_t ($v_t \in \mathbb{R}^{k \times 1}$)。

$$v_t = W \cdot w_t \quad (1)$$

模型输出层是 softmax 函数归一化的条件概率 $p(w_{t+j} | w_t)$, 如式(2)所示。

$$\begin{aligned} w' &= \text{softmax}(W' \cdot v_t + b) \\ p(w_{t+j} | w_t) &= w'_m \end{aligned} \quad (2)$$

其中, $b \in \mathbb{R}^{V \times 1}$, 表示偏置向量; w'_{t+j} 表示列向量 $w' \in \mathbb{R}^{V \times 1}$ 中第 m 个元素, m 与词典顺序有关。

Skip-gram 模型的中心思想是使用中心词 w_t

预测上下文 w_{t+j} , 其训练目标是使得目标函数取最大值, 如式(3)所示。

$$Q = \frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \quad (2)$$

其中, T 表示语料库大小, c 表示上下文窗口长度。

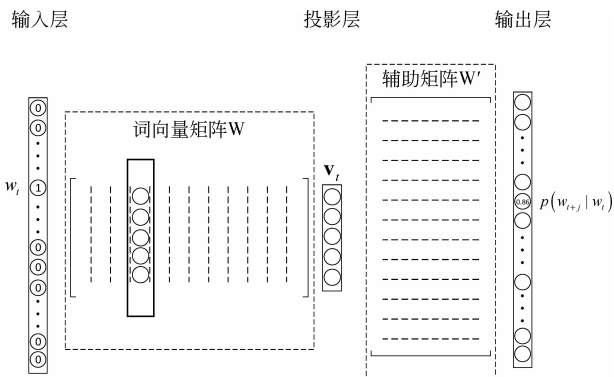


图1 skip-gram 词向量训练模型

通过图 1 可以看出, skip-gram 模型仅通过词语之间位置关系捕捉语义关系。在很多情况下由于语料不是规范的, 例如, 口语语料、微博语料等, 词语之间的相对位置变动较大, 出现较大的训练噪声, 难以训练出所需的高精度词向量。为此, 我们在 skip-gram 方法的基础上, 针对特定应用, 融入词库中词语的类别信息, 提出增强约束词向量并给出了训练方法。

2.2 增强约束词向量训练方法

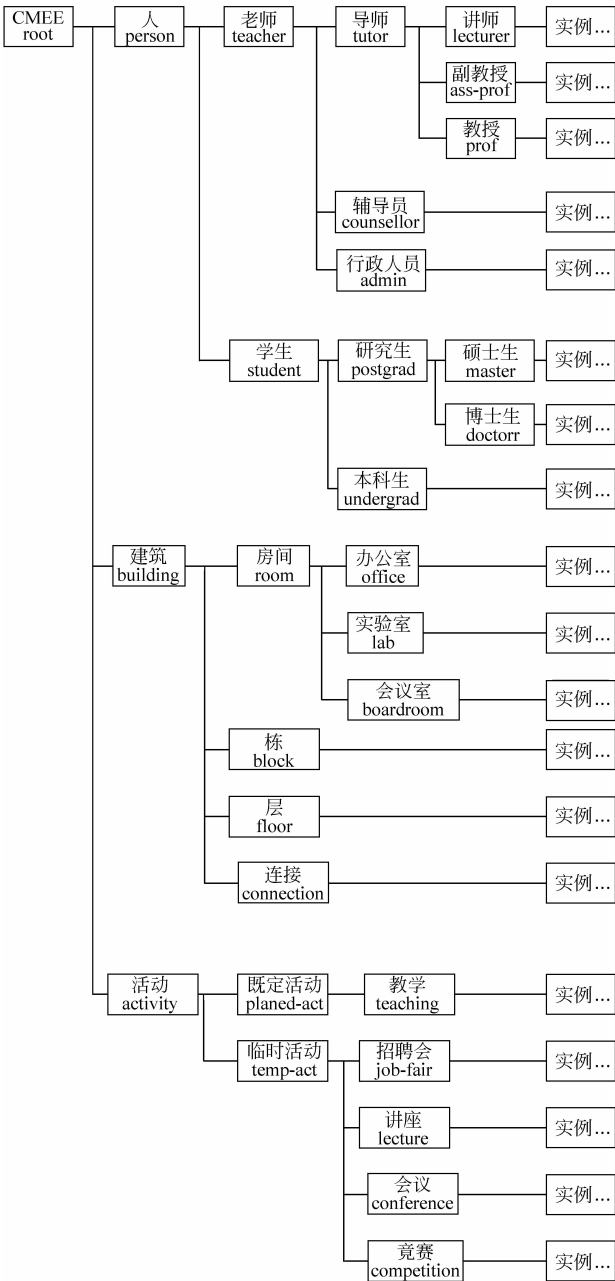
一般词向量在训练中(如 skip-gram, CBOW 等)仅使用词语之间上下文关系这一种约束条件, 词语的其他信息(例如词性)并没有参与到词向量的训练过程中。本文期望通过在词向量的训练过程中, 增加约束条件, 改变词在向量空间的分布, 使其分布更加合理。

本文的目标是限定领域的词向量应用, 通过将特定领域的词进行分类, 建立一个分类本体(ontology)。将这一分类本体看作是该领域的知识, 作为约束条件, 引入到词向量的训练过程中, 期望得到的词向量能够反映这一知识, 形成词向量在空间上更好的分布, 更好地反映该领域的词的语义信息。

在词向量训练过程中, 增加约束的有效方法是对目标函数进行修正, 或者是在训练过程中增加训练目标的多任务训练。图 2 为这种方法的一种实现—constraint-enhanced skip-gram (CE-skip-gram), 在图 1 的 skip-gram 的基础上, 增加了第二任务, 形成了式(4)所示的目标函数。

CE-skip-gram 模型需要对语料 $\{w_1, w_2, \dots,$

之后,利用本体类作为增强约束项,调整词向量的语义表达。



3.2 CE-skip-gram 的构建

CE-skip-gram 方法的关键是利用词语的分类知识约束词向量,本文采用上述基于本体的方式构建关于 CMEE 的领域知识,利用领域知识构造词语的分类标签。由于本应用的特殊性,所有本体词语均是名词,故假设每个词语仅有一个标签。

构造词语分类标签的方法如下:

(1) 构建如图 3 所示的本体知识框架,对本体

上下位关系进行分级,没有父类的本体称为零级本体,有一个父类的本体称为一级本体,其余依此类推。

(2) 利用本体知识框架对词语进行标注,标注粒度根据具体任务而定,从而实现不同粒度的知识表达。本文采用较粗的粒度标注,如果词语出现在本体框架中,且是一级本体,则采用本身的标签;如果是二级及二级以下的本体,采用二级本体作为标签,例如,“讲师”“副教授”“导师”“王 xx”等词的标签都是二级本体“老师”的标签:“teacher”。

(3) 没有出现在本体中的词语统一给定标签“common”。

例如,对于语料“王|教授|的|办公室|在|哪里”,根据上述方法进行知识标注,得到序列“王|nh|教授|teacher|的|common|办公室|room|在|common|哪里|common”。

CE-skip-gram 中的分类器可选常用的 softmax 回归、支持向量机等,本文选用 BP(back propagation)神经网络,输入层是中心词的词向量,输出层选用 softmax 函数进行概率归一化。

由于本文所涉及任务的特殊性(针对 CMEE),无法利用互联网开放的大规模语料,训练词向量数据集全部由实验室众师生与口语对话系统交互所得,共 733 条(表 1)。

表 1 语料示例

序号	语料示例
1	我想找张××教授,怎样走?
2	请问王××在 A 座五楼哪里办公?
3	最近的厕所怎样走?
4	我想去机器人实验室,怎么走?
5	请问机电楼的厕所在哪里?
6	本次会议什么时候开始?
7	李老师的研究生有哪些人?
8	张××的办公室在哪里?
9	本次会议的负责人是谁?
10	今天心理健康知识讲座的主讲人是?

词向量模型训练需要先对训练语料进行分词,我们使用 LTP 平台进行分词,然后根据本体知识框架进行标注。本文增强约束词向量模型训练的基本数据如下:本体知识标签数量为 12;训练数据集共有 733 条语料,词的总数为 599 个;词向量维数 k 设置为 30,约束系数 β 为 0.7,学习率 α 为 0.1,窗口大

小 c 为 2,训练 100 000 步。编程语言采用 Python,词向量模型使用 tensorflow 框架进行编写。

3.3 关于词的歧义性的说明

自然语言处理的难题之一是词语语义歧义性(word sense disambiguation, WSD),因此消歧就成为该领域中的一个重要研究内容。杨陟卓^[21]等采用语言模型优化传统的有监督消歧模型,利用这两种模型的优势,共同推导歧义词的语义;Agirre^[22]等提出了基于词汇知识库的 WSD 算法,实验表明该算法能够更有效地使用 WordNet 图,性能优势明显。

本文中心词的语义类别获取采用的是利用 3.2 节中提到的构造词语分类标签的方法,这里的中心词语义标签实际上是本体标签(根据本体知识库构建标签),例如“teacher”“student”“area”“room”等。在本特定应用中,所有的本体词语均是名词,中心词出现歧义性的情况很少,消歧问题不突出。但随着应用范围的扩大,语义消歧会成为一个问题,需要关注。

4 实验及结果分析

评估词向量有两种方法:①内部任务评价(intrinsic evaluation),②外部任务评价(extrinsic evaluation)。内部任务评价遵循“语义接近的词其词向量也是接近的”原则,通常评价词向量的语义相关性,这种方法需要人工收集近义词表。外部任务评价是在实际任务中对词向量进行评价,例如文本分类任务,通过任务结果来评估词向量,这种与具体任务相结合的评价方法是很有效的^[23]。为验证本文所述 CE-skip-gram 模型的有效性,采用上述两种评价方法。内部任务评价见 4.1 节实验 1,外部任务评价见 4.2 节实验 2、实验 3。

4.1 实验 1: 内部任务评价

本文采用词向量语义相关性实验作为内部任务评价,使用 Word2Vec 研究中惯用的做法:人工标注近义词组,处于同一个近义词组内的词,互为近义词关系。测试数据共 9 组,36 个词,如表 2 所示。

表 2 近义词示例

序号	近义词组
1	老师 讲师 教授 导师

续表

序号	近义词组
2	学生 研究生 硕士 博士 博士生 硕士生
3	办公室 实验室 会议室
4	时间 时候 时长
5	比赛 竞赛 大赛 演讲 招聘会 讲座 宣讲会
6	王 张 李 赵 刘 唐
7	学院 机电院
8	办公室 实验室 会议室
9	哪里 哪

根据训练好的词向量,利用式(4)计算词与词之间的相似度 sim 。每个近义词组可以认为是同一类词,如果与某个词相似性最高的前三个词,与该词所在近义词组存在交集,且相似度 $\text{sim} > 0.65$,就认为该词向量是较为准确的(属于这类);否则,认为是不准确的(属于其他类)。这样评价词向量语义相关性就可以看作是多分类任务,评估指标选用精确率(P)、召回率(R)、综合二者的 F_1 值以及准确率(ACC),计算方法如式(7)~式(10)所示。实验结果如表 3、表 4 所示。

$$\text{sim}(v_1, v_2) = \cos(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1| \times |v_2|} \quad (6)$$

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (9)$$

$$\text{ACC} = \frac{\text{正确预测的数目}}{\text{预测的总数目}} \quad (10)$$

其中 TP 表示将正类预测为正类数;FN 表示将正类预测为负类数;FP 表示将负类预测为正类数;TN 表示将负类预测为负类数。对于多分类任务,把每个类别单独视为“正”,所有其他类别视为“负”。

根据表 3~4 的实验结果可知,相较于 skip-gram 模型,CE-skip-gram 模型所得到的词向量准确率更高,近义词词向量聚集更紧密。这个结果符合我们的预期。口语语料大多不规整,词语之间相对位置多变,skip-gram 模型仅靠词语之间上下文关系约束词向量矩阵,出现较大训练噪声,所得词向量表达能力有限,而 CE-skip-gram 模型,在利用上下文关系作为约束条件的基础上,增加新的约束条件,利用词语分类信息约束词向量矩阵,减小了噪声

表 3 skip-gram 模型与 CE-skip-gram 模型对比

	skip-gram				CE-skip-gram			
近义词组	<i>P</i>	<i>R</i>	<i>F</i> ₁	ACC	<i>P</i>	<i>R</i>	<i>F</i> ₁	ACC
词组 1	1.0	0.5	0.667	——	1.0	1.0	1.0	——
词组 2	1.0	1.0	1.0	——	1.0	1.0	1.0	——
词组 3	1.0	0.333	0.5	——	1.0	1.0	1.0	——
词组 4	1.0	0.5	0.667	——	1.0	0.75	0.857	——
词组 5	1.0	0.833	0.909	——	1.0	0.833	0.909	——
词组 6	1.0	1.0	1.0	——	1.0	1.0	1.0	——
词组 7	1.0	1.0	1.0	——	1.0	1.0	1.0	——
词组 8	1.0	0.333	0.5	——	0.667	0.667	0.667	——
词组 9	0	0	0	——	0.667	1.0	0.667	——
总体	0.944	0.703	0.806	0.638	0.972	0.919	0.945	0.861

表 4 测试结果的部分细节

原词	方法	排前三的相似词及相似性					
硕士	skip-gram	博士	0.79	博士生	0.76	硕士生	0.75
	CE-skip-gram	博士生	0.95	博士	0.94	硕士生	0.91
招聘会	skip-gram	宣讲会	0.76	讲座	0.66	党员大会	0.63
	CE-skip-gram	会议	0.83	讲座	0.72	党员大会	0.68
老师	skip-gram	饮水机	0.63	负责人	0.59	主持人	0.57
	CE-skip-gram	副教授	0.85	教授	0.83	导师	0.70
办公室	skip-gram	讲座	0.79	饮水机	0.76	负责人	0.75
	CE-skip-gram	教室	0.85	会议室	0.84	实验室	0.82
刘	skip-gram	李	0.85	张	0.79	孔	0.78
	CE-skip-gram	张	0.90	李	0.89	孔	0.87

的影响,缩小同类词之间的“距离”,使同类词语的词向量分布更加紧密,所得词向量更加准确。

4.2 外部任务评价

本文所设计的信息查询系统中,语义理解模块是其关键组成部分(其中的信息查询系统未在本文中详细介绍)。该模块需要对语句进行解析,包括两个子任务:意图理解(intent understanding)与槽填充(slot filling)。“意图理解”是对语句的整体按意图进行分类,“槽”指的是每一种意图的语句经过结构化处理的每一个要素的位置。

根据本文的应用场景,将用户对话意图分为以下 7 种:Query_student,Query_location,Query_activity,Introduce,Query_teacher,Query_org,Con-

firm,见表 5;槽分为以下 20 种:teacher,where,name,who,nh,area,activity,p-prop,room,a-prop,org,list,person,location,research,att,

表 5 语句意图示例

序号	意图	语料	说明
1	Introduce	介绍一下朱××老师	请求介绍
2	Query_activity	本次会议的主题是	询问活动
3	Query_location	一楼的厕所怎么走	导航相关
4	Query_org	机电学院有多少老师	询问机构
5	Query_teacher	王××教授研究方向是	询问老师
6	Query_student	张××的导师是谁	询问学生
7	Confirm	请问 A513 是办公室吗	请求确认

o, student, department, thing, 语句结构化示例见表 6。例如, 语句“王××|的|研究生|有|哪些|人”, 经过语义理解模块, 输出意图为“Query_teacher”类, 语句的相应要素为“name|o| student|o| which|o”。下一步需要根据所得意图与要素, 生成知识库查询语句。

表 6 语句结构化信息示例

序号	语句						
1	王	老师	在	哪里	?		
	nh	teacher	o	where	o		
2	张	教授	的	办公室	在	哪里	?
	nh	teacher	o	room	o	where	o
3	本次	讲座	的	主讲人	是	谁	?
	att	activity	o	a-prop	o	o	o
4	请问	招聘会	在	哪里	举行	?	
	o	activity	o	where	o	o	
5	我	想去	C210	怎样	走	?	
	o	o	room	o	o	o	

该系统的知识库通过本体建立起来, 是基于谓词逻辑的 RDF 描述文档。知识库的查询也是使用基于谓词逻辑的查询语句。本系统使用 SWI-PROLOG 语言, 但知识库及 Prolog 的处理过程不在本文的介绍范围之内。

上述例句生成的查询语句可以表示为(A, is_a_student_of, 王××)和(A, is_a, 研究生)。其中两个谓词分别为 is_a_student_of 和 is_a, 未知变量为 A。通过查询知识库, 可以得到结果 A。文本处理流程如图 4 所示。

本文所采用的外部任务评价就是利用词向量实现上述语义理解任务, 通过评价任务来评价词向量。外部任务评价包括两部分, 其一是意图识别实验(实验 2), 其二是槽填充实验(实验 3)。将数据集按照 2: 8 比例分成测试集(147 条)和训练集(586 条), 模型评估指标选用精确率(P)与召回率(R)以及综合二者的 F₁ 值。

4.2.1 实验 2: 意图识别实验

对话意图识别是判别用户的意图(目的), 是一种分类问题。我们分别将随机数词向量、skip-gram 词向量与 CE-skip-gram 词向量作为分类器的输入, 评价分类器的效果。使用的分类器分别为 K 近邻(KNN)、支持向量机(SVM)、卷积神经网络(CNN)

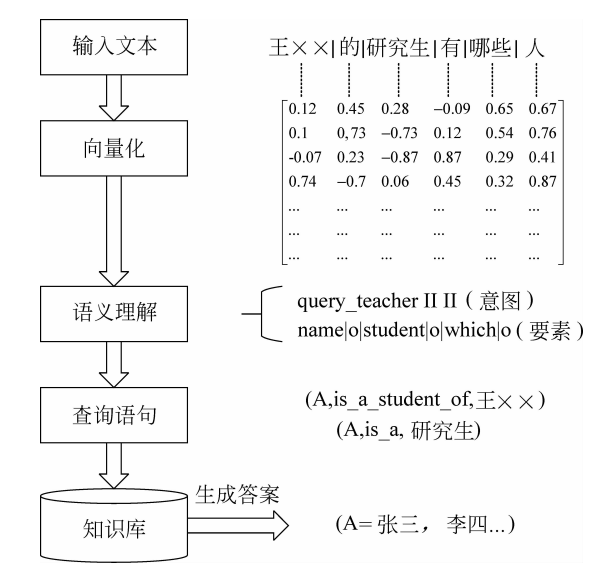


图 4 文本处理流程图

和循环神经网络(RNN)。
KNN、SVM 模型的输入是词向量的连接 $w_{input} = [w_{t-i} \oplus \dots \oplus w_t \oplus \dots \oplus w_{t+j}]$, 构成的一个长向量 $w_{input} \in \mathbb{R}^{m \times 1}, m = l \times k$, 其中 l 表示语句的长度, k 表示词向量的维度, \oplus 表示连接操作。KNN、SVM 模型使用 sklearn 机器学习工具包实现, 模型参数设置见表 7。

表 7 分类器模型参数设置(1)

KNN		SVM	
参数	值	参数	值
n_neighbors	5	C	33
weights	uniform	kernel	rbf
leaf_size	1.0	degree	3
p	2	shrinking	True
metric	minkowski	probability	False
n_jobs	1	tol	0.001
—	—	cache_size	200

CNN 模型^[24]如图 5 所示, 是一个 5 层的神经网络。CNN 模型的输入是词向量的堆叠 $w_{input} = [w_{t-i}; \dots; w_t; \dots; w_{t+j}]$, 构成一个矩阵 $w_{input} \in \mathbb{R}^{k \times l}$; 第一层为卷积层, 采用宽度为 3、4、5 的三种卷积窗口, 每种窗口有 8 个卷积核用于特征提取; 池化层采用最大池化操作; 在拼接层将所得特征全部拼接; 经过全连接层后进行 softmax 操作, 得到分类结果。模型参数设置见表 8。

RNN 模型^[25]是按照时间顺序输入语句中每个

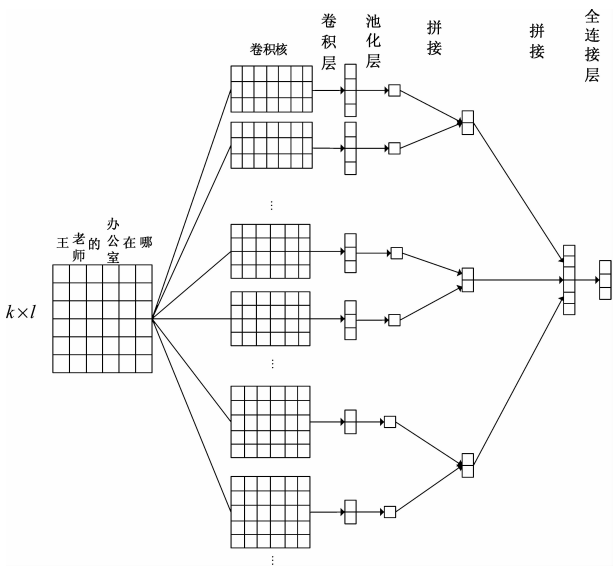


图 5 CNN 模型示意图

词的词向量 w_i ，如图 6 所示。RNN 神经网络单元采用双向长短期记忆神经网络 (Bi-LSTM)，前向 LSTM 读取序列的正向信息，得到前向状态 fh_i ，反向 LSTM 读取序列的反向信息，得到反向状态 bh_i ，状态 $h_i = [fh_i \oplus bh_i]$ 。最终状态 h_{last} 接全连接层，进行 softmax 操作，得到分类结果。模型参数设置见表 8。

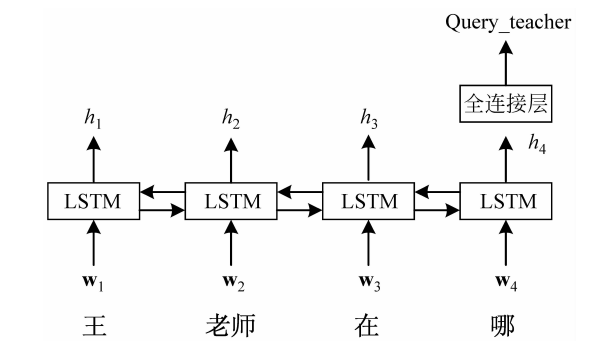


图 6 RNN 模型示意图

表 8 分类器模型参数设置(2)

CNN		RNN	
参数	值	参数	值
fliter_size	3,4,5	input_steps	30
num_filter	8	hidden_size	100
dropout_keep_prob	0.7	layer_num	2
batch_size	8	batch_size	8
num_epoch	200	num_epoch	50

实验结果分析总结为以下两点：

(1) 从整体上看，四种机器学习模型均在使用 CE-skip-gram 词向量作为特征向量时，效果最优。原因在于随机数中未包含任何语义信息，skip-gram 虽然可以提取到语义信息，但所提取到的语义信息不够准确，而 CE-skip-gram 可以提取到准确的语义信息，在 skip-gram 的基础上使机器学习模型学习效果进一步提升。

(2) 观察表 9~12 的实验结果，使用 skip-gram 词向量相比于使用随机数词向量，会给模型带来较大的提升，尤其是 CNN 模型， F_1 值提升了 25.8%。但是这种提升效果在实验 4 中表现并不明显，原因在于 RNN 模型本身是序列模型，文本信息隐藏在序列中，模型不完全依赖词向量提供的语义信息，而 KNN、SVM、CNN 模型几乎不考虑文本的顺序，丢失了文本的序列信息，只能通过词向量获取语义信息，所以词向量构造的好坏直接影响模型的结果。

表 9 实验结果 1: KNN 模型实验结果

	P	R	F_1
随机数	0.745	0.657	0.698
skip-gram	0.787	0.701	0.741
CE-skip-gram	0.791	0.73	0.759

表 10 实验结果 2: SVM 模型实验结果

	P	R	F_1
随机数	0.336	0.415	0.372
skip-gram	0.472	0.522	0.496
CE-skip-gram	0.473	0.540	0.504

表 11 实验结果 3: CNN 模型实验结果

	P	R	F_1
随机数	0.587	0.703	0.638
skip-gram	0.903	0.890	0.896
CE-skip-gram	0.940	0.891	0.914

表 12 实验结果 4: RNN 模型实验结果

	P	R	F_1
随机数	0.931	0.879	0.903
skip-gram	0.937	0.914	0.925
CE-skip-gram	0.963	0.927	0.945

4.2.2 实验 3：槽填充实验

槽填充是指从用户对话中提取到与任务相关的关键信息。例如，在本文所涉及的对话任务中，老师、学生、活动地点等词就是关键的槽位信息。通过提取到的槽位信息，生成查询知识库的语句，得到所需的答案。槽填充任务实质是序列标注问题，本文采用的是基于注意力机制的编码-解码(encoder-decoder)模型^[23]。

encoder-decoder 模型是一种 Seq2Seq (sequence to sequence)模型，如图 7 所示。在编码端，使用Bi-LSTM神经网络，前向 LSTM 读取序列的正向信息，得到前向状态 fh_i ，反向 LSTM 读取序列的反向信息，得到反向状态 bh_i ，在第 i 步的状态 $h_i = [fh_i \oplus bh_i]$ 。在解码端，使用单向 LSTM 神经网络，采用注意力机制^[26]，对每个状态 h_i 进行解码，得到序列标注。模型参数设置见表 13。

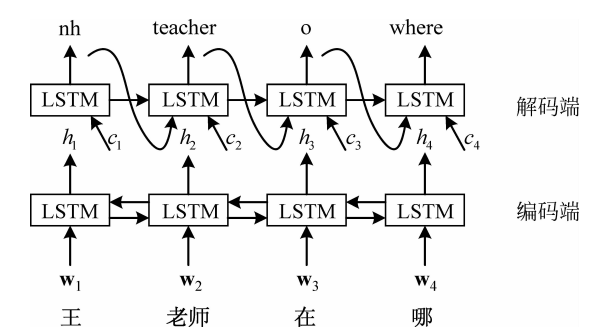


图 7 encoder-decoder 模型示意图

表 13 encoder-decoder 模型参数设置

参数	值
input_steps	30
hidden_size	100
layer_num	2
batch_size	8
num_epoch	50

观察表 14 的实验结果可以发现，①使用 CE-skip-gram 词向量的实验结果最优，使用 skip-gram 词向量与随机数词向量的结果大体一致。原因在于序列模型(Seq2Seq 模型)主要利用文本的序列信息，而不是语义信息，所以即使是不包含任何语义信息的随机数词向量作为输入特征，模型也能达到可观的效果；②在同样的序列信息基础上，使用语义精度较高的 CE-skip-gram 词向量，可以进一步提升模型的效果。

表 14 实验结果 5：槽填充实验结果

	P	R	F_1
随机数	0.972	0.970	0.971
skip-gram	0.972	0.974	0.973
CE-skip-gram	0.985	0.986	0.986

5 结论

纵观自然语言处理研究的历史，理性主义方法与经验主义方法此消彼长。尽管近年来经验主义方法利用大规模语料，取得了一定的成功，但是它在理论的完备性上存在不足。要想彻底解决自然语言处理问题，必须将这两种研究方法结合起来。基于这种思想，本文提出了一种词向量训练方法——增强约束词向量模型。在利用词语上下文关系作为约束的基础上，将任务相关的知识作为增强约束项，干预词向量的生成。针对具体任务(限定场合的对话系统)，我们首先利用本体表达领域知识，之后根据领域知识对词语进行标注，通过多任务学习的机制，将预测中心词的知识标签作为附加任务，对词向量矩阵加以约束，从而将知识信息引入词向量中。这样，词向量中蕴含的语义信息在人工知识的帮助下得以修正，使表达更加精确。采用内部任务和外部任务两种方法对词向量进行评估与对比，结果表明本文提出的增强约束词向量在表达词的语义信息方面更加准确，将其应用于特定场合的对话系统也得到了更好的意图理解效果，对提高自然语言对话的自然流畅性有较大的帮助。

本文的解决思路，用于提升领域任务型对话系统的语义理解与对话的自然流畅性具有一定的普适意义，对大型的商场、医院、地下车场、旅游景点等场合的口语导引系统等具有一定的借鉴意义。本文对于中心词的类别处理是根据知识库人工给出的，没有实现自动类别处理。为提高效率可进一步研究自动化获取领域知识的方法，增加本方法的适用性。

参考文献

[1] De Saussure F. Course in general linguistics, etc[M]. McGraw-Hill Book Company, 1966.

[2] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25(2-3): 259-284.

- [3] Nightingale C, Myers D J, Linggard R. Introduction neural networks for vision, speech and natural language[M]. Neural networks for vision, speech and natural language. Springer, Dordrecht, 1992: 1-4.
- [4] Jurafsky D. A probabilistic model of lexical and syntactic access and disambiguation[J]. Cognitive Science, 1996, 20(2): 137-194.
- [5] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013: 9-18.
- [6] Xu P, Sarikaya R. Convolutional neural network based triangular CRF for joint intent detection and slot filling [C]//Proceedings of the Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 78-83.
- [7] Zhang X, Wang H. A joint model of intent determination and slot filling for spoken language understanding [C]//Proceedings of the IJCAI, 2016: 2993-2999.
- [8] Harris Z S. Distributional structure[J]. Word, 1954, 10(2-3): 146-162.
- [9] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(Feb): 1137-1155.
- [10] 冯艳红, 于红, 孙庚, 等. 基于词向量和条件随机场的领域术语识别方法[J]. 计算机应用, 2016, 36(11): 3146-3151.
- [11] Liao C, Feng C, Yang S, et al. Topic-related Chinese message sentiment analysis[J]. Neurocomputing, 2016, 210(C): 237-246.
- [12] Bian J, Gao B, Liu T Y. Knowledge-powered deep learning for word embedding[M]. Machine Learning and Knowledge Discovery in Databases, 2014: 132-148.
- [13] 杜慧, 徐学可, 伍大勇, 等. 基于情感词向量的微博情感分类[J]. 中文信息学报, 2017, 31(3): 170-176.
- [14] Hu B, Tang B, Chen Q, et al. A novel word embedding learning model using the dissociation between nouns and verbs[J]. Neurocomputing, 2016, 171: 1108-1117.
- [15] Liu M, Lang B, Gu Z, et al. Measuring similarity of academic articles with semantic profile and joint word embedding[J]. Tsinghua Science and Technology, 2017, 22(6): 619-632.
- [16] Chen Z, He Z, Liu X, et al. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases[J]. BMC Medical Informatics and Decision Making, 2018, 18(2): 65.
- [17] Taghipour K, Ng H T. Semi-supervised word sense disambiguation using word embeddings in general and specific domains[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015: 314-323.
- [18] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [19] 张梅, 郝佳, 阎艳, 等. 基于本体的知识建模技术[J]. 北京理工大学学报, 2010, 30(12): 1405-1408.
- [20] Neches R, Fikes R E, Finin T, et al. Enabling technology for knowledge sharing [J]. AI Magazine, 1991, 12(3): 36-56.
- [21] 杨陟卓, 黄海燕. 基于语言模型的有监督词义消歧模型优化研究[J]. 中文信息学报, 2014, 28(1): 19-25.
- [22] Agirre E, López de Lacalle O, Soroa A. Random walks for knowledge-based word sense disambiguation[J]. Computational Linguistics, 2014, 40(1): 57-84.
- [23] Li Y, Yang T. Word embedding for understanding natural language: A survey[M]. Guide to Big Data Applications. Springer International Publishing, 2018: 83-104.
- [24] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [25] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling [J]. arXiv preprint arXiv:1609.01454, 2016.
- [26] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv:1409.0473, 2014.



王恒升(1963—), 博士, 博士生导师, 主要研究领域为智能机器人和机电一体化。

E-mail: whsheng@csu.edu.cn



任晋(1990—), 博士, 主要研究领域为智能机器人、知识表达与推理、人机交互。

E-mail: 591726822@qq.com



刘通(1994—), 通信作者, 硕士, 主要研究领域为自然语言处理与人机交互。

E-mail: liutong9404@qq.com