基于聚类和相似度计算的陆空通话词向量评估

向 倩
(中国民航大学 空中交通管理学院,天津 300300)

摘 要:无线电陆空通话是管制员与飞行员进行话音通信的方式,对航空器运行有着重要作用。在陆空通话用语的处理中,词向量是充分表征词汇语义的有效表现形式。为保证管制员飞行员人机对话系统词向量输入质量,提出了基于 K-Means 概念分类和基于孪生网络句子相似度计算的词向量评估方法。概念分类实验分析了单词依托向量映射到手工分类词典的准确率,结果显示准确率平均值达 80.2%,浅层证明词向量具备表征语义区分单词的能力,符合空管指令分类明显的特征。句子相似度计算利用基于 Siamese 网络的模型计算了空管指令对的相似度值,该模型与基于 wordnet 层级距离、基于编辑距离方法的相似判断准确率分别为 93.6%、65.8%、43.7%,前者远超其他两种方法,深层证明词向量能充分捕获词汇语义,满足对话系统对词向量质量的输入需求。

关键词:陆空通话;词向量;概念分类;句子相似度;孪生网络
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2020)09-0137-06
doi:10.3969/j.issn.1673-629X.2020.09.025

Word Embeddings Evaluation Based on Clustering and Similarity Computing in Radiotelephony Communications

XIANG Qian
(School of Air Traffic Management, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Radiotelephony communications is a means of voice communication between controllers and pilots, which plays an important role in the aircraft operations. In the processing of radiotelephony communications, word embeddings is an effective representation to capture lexical semantics. To ensure the quality of word embeddings inputting into the human-machine dialogue system between controllers and pilots, an evaluation method combining concept categorization is proposed based on K-Means and sentence similarity computing based on Siamese network. In the experiment of concept categorization, the accuracy of words mapping to manual dictionary is analyzed, which is up to 80.2%. It is proved that the word embeddings has the ability of representing semantics and distinguishing words, which is consistent with the classification feature of radiotelephony communications. The similarity of two instructions is calculated by Siamese-based network model in sentence similarity computing, and the accuracy of this model is 93.6% which highly exceeds the models based on wordnet hierarchy distance (65.8%) and edit distance (43.7%). The result shows that word embeddings can fully capture lexical semantics and meet the input requirements of the dialogue system for the word embeddings quality.

Key words: radiotelephony communications; word embeddings; concept categorization; sentence similarity; Siamese network

0 引 言

陆空通话用于管制员与飞行员交互空中交通动态信息,是保障航空安全和效率的最基础手段,是一种全球通用的管制员和飞行员英文对话标准。为加强管制员培训效果,满足非英语国家陆空通话用语规范和发音标准的需求,提出管制员飞行员人机对话系统,利用计算机响应管制指令替代机长席位。对话系统分为语音识别模块和语言理解模块。语音识别与语音合成技术[1-2]已广泛应用于陆空通话领域,而直到最近几年,自然语言理解研究才得到了初步发展。2017 年,卢薇冰[3]和路玉君[4]基于改进 CNN 和 RNN 模型利用词向量对陆空通话语义相似度进行比较,辅助计算机判断复诵过程中的失误。

2013 年,词向量被证明能捕获更复杂的语言属性,最大限度保留单词语义、结构信息[5]。人机对话系统语言理解模块基于神经网络,以对话文本为网络实际输入,文本向量为本质输入,高质量的词向量能赋予神经网络更多学习信息,因此除优化系统网络模型外,

收稿日期:2019-11-11 **修回日期**:2020-03-17
基金项目:国家自然科学基金(71801215)
作者简介:向 倩(1993-),女,硕士研究生,研究方向为自然语言处理、陆空通话。

需要对词向量进行评价以保证源头输入向量的质量。

早期词向量内部评估通过直接测量语义相关性和几何相关性来测量词向量的质量,包括相似性计算、类比、分类等方法[6]。后来的学者更加关注词向量在下游任务的表现,使用词向量作为下游任务的输入特性,并度量特定于该任务的性能指标的变化称为外部评估方法。Schnabel[7]、Anna[8]等人结合了传统内部评估方法和诸如名词短语分块、命名实体识别、情绪分类、推理任务等外部评估方法。Tulkens[9]提出了利用词向量将荷兰语方言文本映射到文本原始分类区域,借此衡量词向量的相似分类特性。

对于低资源语言如陆空通话用语,由于缺乏前人标注文件和工具材料,需要制定自身的评估方法和标准。词向量的首要功能是表征语义,对词向量质量评估即对词向量表征语义的能力进行评估。考虑到下游管制员飞行员人机对话系统的任务,制定了一个基于 K-Means 的概念分类和基于 Siamese 网络句子相似度计算的陆空通话词向量评估方法。首先建立陆空通话数据集,借助 word2vec 模型生成词向量;其次利用概念分类的方法,通过比较词向量分类和人工分类词典的差异来证明词向量表征语义区分单词的功能;最后建立陆空通话指令比较集,通过词向量来比较指令相似度,利用判断准确率来进一步证明词向量表征语义的功能。

1 语义表示方法

自然语言理解的发展得益于语义表达技术的发展,早期人类知识被表示为知识库的形式;随着计算机技术的更新,以自然文本为输入,高维稀疏向量为输出的传统语义表达方法开始盛行[10];然而高维稀疏的语义表达方法无法有效地表达出词语之间的相似度信息,1986 年 Hinton[11]提出了词的分布式表示,能通过刻画词的多重属性更高效表示词义和语言结构等信息,在形式上表示为低维连续的向量。以管制员指挥国航 1421 航班调整航向的指令为例,展示了两种语义表示方法的具体形式,如表 1 所示。

表 1 语义表示方式示例

单词	One-Hot	Word embedding
CCA1421	(1,0,0,0)	(1.184 499 6,-0.504 033 15,…,-0.168 002 5)
fly	(0,1,0,0)	((-1.849 144,0.578 670 56,…,-1.139 345)
heading	(0,0,1,0)	(-0.586 067 26,4.319 369,…,-0.712 084 7)
210	(0,0,0,1)	(0.796 223 3,-1.966 892 7,…,-0.355 712 5)

注:“CCA1421,fly heading 210.”语义表示方法。

每个单词以几十上百维的向量形式表示,涵盖了语义、语法、上下文关系等多种特征。词向量将单词映射到向量空间里,通过计算单词的“距离”信息来捕捉它们之间存在的句法(结构)和语义(语义)等相关关系。

2 陆空通话词向量训练

2.1 陆空通话语言结构分析

陆空通话语料数据集来源于飞行进离场阶段真实通话录音文件。进场阶段是指航路飞行航空器下降对准跑道的过程,离场阶段是指离场航空器加入航路飞行的阶段。该阶段涉及到的空管指令主要包括:高度、速度、航向、进离场程序。其对话结构如下:

对方呼号+通话内容

复述通话内容+己方呼号

示例:

C:CCA1421,Dongfang Approach,radar contact.

P:Dongfang Approach,CCA1421.

C:CCA1421,turn right heading 110 for spacing.

P:Right heading 110,CCA1421.

管制员用语分为许可类、指令类、限制类、报告类、证实类和信息类,飞行员用语分为请求类、状态报告类、复讼类。

陆空通话标准用语具有以下几个特点:语法结构单一,指令长度适中,指令重复率高,词汇量有限,属于小型语料库。分析陆空通话语言结构及词向量评估方法可得出:

(1)对话多为祈使句和陈述句,不具有主观情绪色彩,因此不能用情感分类来评估陆空通话词向量。

(2)语料库词汇较少,指令之间存在较弱的上下文语境关系,单词之间缺乏有效的类比关系,故不能用类比(关系识别)来评估。

(3)指令主语一般为航空器呼号,选择偏好用于判断句子语义和动宾等逻辑信息,同样也不适合该方法。

多个传统词向量内部评估标准均不适用于陆空通话词向量的评估。结合管制员飞行员人机对话系统需求和词向量内部、外部评估标准,利用概念分类和句子相似度计算来评价陆空通话词向量。

2.2　词向量生成训练

根据上述陆空通话呼叫结构形式和终端区信息,建立指令类-复讼类常规陆空通话数据集,共计 360 个单词,3 167 641 条指令-回答语句对,涉及 5 架航空器、1 家管制单位。

目前应用最为广泛的词向量训练方法有 word2vec、Glove,经过众多研究显示 word2vec 在大部分测评指标优于 Glove。word2vec 可利用 CBOW 和 Skip-gram 两种方法产生词向量,CBOW 是输入已知上下文,输出对当前单词预测的模型,Skip-Gram 是推测当前单词上下文单词的模型,模型网络结构见图 1。

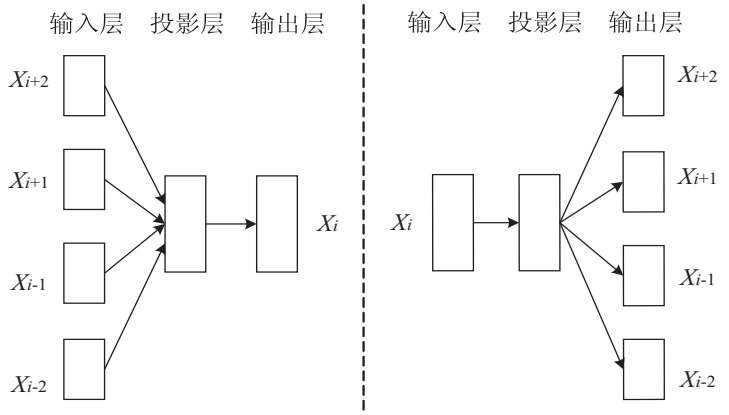


图 1　CBOW 网络结构、Skip-gram 网络结构

简而言之,word2vec 模型其实是一个由输入层、隐藏层、输出层组成的简单神经网络,隐藏层为线性的单元。该模型以 One-Hot 向量为输入,经过训练之后,使用输入层和隐藏层之间的连接权重矩阵表示单词之间的关系,输出层与输入层具有相同维度。

(1)CBOW 模型。

CBOW 模型又称连续词袋模型,以某中心词临近的上下文单词所对应的词向量为输入,输出该特定中心词的词向量。

模型训练过程为:首先经输入层输入中心词上下文单词的 one-hot 向量,设其空间维度为 D,上下文单词个数为 C;C 个 one-hot 向量分别乘以共享的输入权重矩阵 $\boldsymbol{M}_{D\times N}$,之后将所得向量求均值作为隐藏层向量,维度为 $1\times N$;隐藏层向量乘以输出权重矩阵 $\boldsymbol{M}'_{D\times N}$ 后,再经激活函数处理得到中间词的概率分布情况,概率最大的索引所指示的单词即为预测出的中间词;将中间词与实际单词对应关系之间进行比较,不断减小误差更新权重矩阵,直到损失值最小。损失函数一般采用交叉熵代价函数,$\boldsymbol{M}$ 和 $\boldsymbol{M}'$ 权重矩阵的计算一般采用梯度下降法。训练结束,输入的每个单词与权重矩阵 $\boldsymbol{M}$ 相乘所得向量即为目标词向量。

(2)Skip-gram 模型。

Skip-gram 颠倒了 CBOW 的输入输出关系,即已知当前单词,预测其上下文单词。不根据上下文单词来猜测目标单词,而是推测当前单词可能的前后单词。该模型输入为某一中心词的词向量,而输出则是该中心词对应的上下文词向量。

建立一个可以沿文本滑动的时间窗,窗口大小 N 表示窗里含特定词在内的单词数目,利用该滑动窗就能统计出每个单词可能出现的上下文单词;为加快训练速度,将预测相邻单词这一任务改变为提取输入与输出单词的模型,并输出一个表明它们是否是邻居的分数(0 表示“不是邻居”,1 表示“邻居”)。这个简单的变换将需要的模型从神经网络改为逻辑回归模型,因此更简单,计算速度更快。同时为避免所有例子都是邻居即准确率为 100% 时而产生低质量词向量,可在数据集中引入不是邻居单词样本作为负样本,为这些样本返回 0,并随机填充输出单词;最后训练神经网络模型,减小损失值,不断更新模型参数用以表示单词之间的关系。

CBOW 在小型数据库中表现更佳,而 Skip-gram 多用于大型语料库,陆空通话语料库为小型语料库,更适用前者。词向量生成训练选择 gensim 库中的 word2vec 模块,在陆空通话数据集上进行训练产生词向量,窗口大小设置为 5,维度设置为 300。

3　陆空通话词向量评估

传统词向量评估基于更高得分向量必然会改善下游任务结果的假设,往往根据得分来比选不同方法产生的词向量。然而该假设有时并不成立,不同的自然语言处理任务可能依赖于词向量的不同语言特征,不能将这些评估分数用作向量质量的绝对评估标准。词向量作为无监督技术的产物缺乏目标值比较,如果不参考下游任务的性能,不能较客观地对其质量进行评估,因此针对特殊任务需要一套绝对的指标来评估词向量。

陆空通话指令分类明显,内部评估任务设定为概念分类实验,根据词向量将具有相同相似属性的单词聚类,浅层分析词向量词义和词性的表征效果[12]。外部评估任务设定为指令相似度计算实验,通过对比基于词向量的句子相似度计算方法和基于 wordnet、基于编辑距离的方法来佐证词向量表征语义的功能。

3.1 概念分类

手工制作陆空通话词典并标注分类,利用词向量将其所对应文本映射到文本的手工分类区域,通过计算映射准确率,判断是否可以在没有任何监督的情况下创建与手工制作资源等价的资源。

手工词典按指令类别分为 4 大类和 19 小类。大类:速度、高度、航向、其他进离场指令。小类:航向动作、航向数据、高度动作、高度数据、速度动作、速度数据、动作原因、方位、跑道、航路点、航空器呼号、进离场程序、管制单位、管制单位频率、距离数据、气压数据、应答机编码、介词、其他单词。

利用 K-means 将词向量聚类为 10 类。聚类数据 i 映射到原始分类中的数据个数和所占比例见公式 1。各类别映射准确率见表 2。

$$p=\begin{cases}p(i\mid m) & \text{if count}(i) > 30\% * \text{count}(m)\\ p(i\mid n) & \text{otherwise}\end{cases} \tag{1}$$

其中,i 为聚类类别,m 为大类类别($\leqslant 4$),n 为小类类别($\leqslant 19$),count() 为数据个数。

表 2 聚类数据映射到原始分类中数据的比例

聚类编号	词典编号	比例/%	聚类编号	词典编号	比例/%
1	小-距离	100	6	大-高度	42.3
2	小-位置	100	7	大-航向	54.6
3	小-高度	100	8	—	—
4	小-航向	64.7	9	—	—
5	小-程序	100	10	—	—
				平均值:	80.2

注:数据过少的分组忽略不计。

表 2 聚类数据的映射结果显示,当聚类数据较少时映射准确率较高,普遍能达到 90% 以上,分类特征明显。而当聚类数据较多时,由于单词涉及的航行动作多且复杂,映射准确率较低,不到 60%。该结果从浅层证明了陆空通话词向量具有一定分类特性,符合陆空通话指令分类明显的特征。

利用 t-SNE(t-distributed stochastic neighbor embedding)机器学习算法将 300 维词向量降至 2 维,图 2 可视化了词向量在二维空间的分布和词向量聚类结果,不同灰度代表不同单词分类。其中词向量的空间距离不代表向量的实际距离。

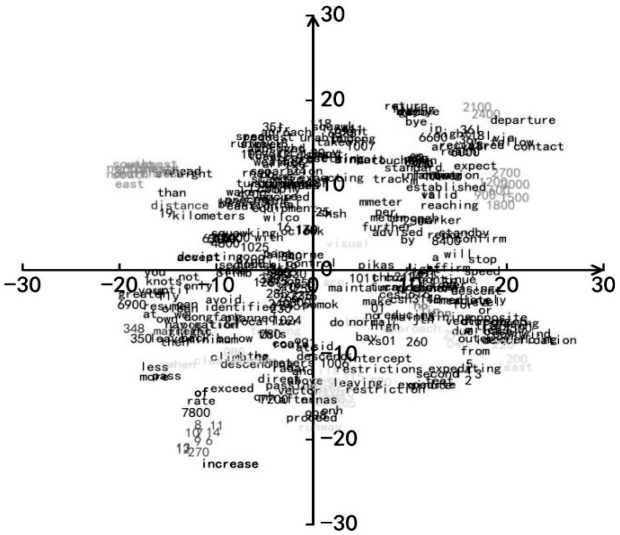


图 2 陆空通话词向量聚类二维平面图

3.2 句子相似度计算

利用基于 Siamese 网络的句子相似度计算模型来判断两句陆空通话指令的相似程度,其结构如图 3 所示。通过比较该方法与其他计算方法的判断准确率来评价词向量。句子相似度计算的其他方法主要有基于编辑距离的方法、基于 wordnet 层级距离的方法,及其

组合方法。

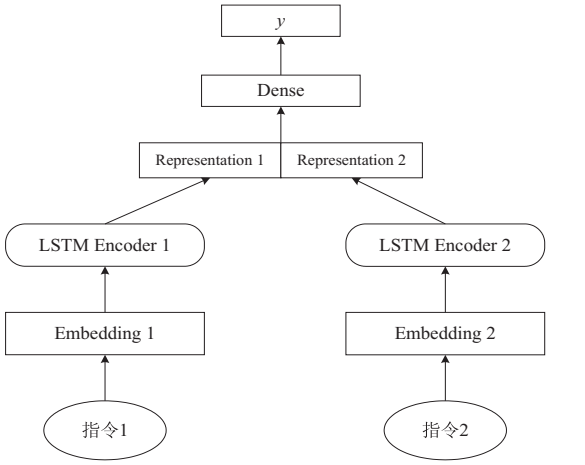


图 3　句子相似度计算模型结构

基于编辑距离的方法是指计算两个句子之间,由一句话转成另一句话所需的最少编辑操作次数,次数越多,说明它们越不同,多以单词共现程度来衡量两句话相似度。这种方法从单词和语句表面结构出发,弱化了同义词的语义关系,可视为基于语言结构的度量方法[13]。

wordnet 词典详细定义了每个单位的词性和词义,利用单词上下位关系构成分类树,基于 wordnet 层级距离的方法将分类树中的路径作为相似度计算的参数。这种方法从语句深层词义出发,可视为基于单词词义的度量方法,容易忽略掉反义词包含的可用上下文相关性[14]。对于字面不相似语义相似、语义不相似句子结构相似的场景需要更复杂的模型来捕捉语义和结构信息。

(1)网络结构。

Siamese 网络是一种神经网络框架,利用 2 个共享权值的网络学习一对输入数据的差异,能同时考虑单词词义和语言结构[15]。具体使用 LSTM 来构建句子相似度计算模型,该模型由输入层、嵌入层、LSTM 层、全连接层和输出层 5 部分组成。LSTM 读取表示每个输入句子的词向量,它的最终隐藏状态即为每个句子的向量表示,这些句向量由词向量构成,它们之间的相似性被用作语义相似性的预测。该方法依赖于预先训练好的词向量作为 LSTM 输入,因此它将受益于词向量质量的提升。损失函数选择交叉熵损失函数 binary_cross entropy,当 x_1 和 x_2 相等时 loss=0,否则 loss 为一个正数,概率相差越大,loss 越大。

$$\begin{aligned} \text{loss} &= -\sum_{i=1}^{n} x_{i2}\log x_{i1} + (1 - x_{i2})\log(1 - x_{i2}) \\ \frac{\partial \text{loss}}{\partial x} &= -\sum_{i=1}^{n} \frac{x_{i2}}{x_{i1}} - \frac{1 - x_{i2}}{1 - x_{i1}} \end{aligned} \tag{2}$$

其中,x_{i1} 表示第 i 个样本第 1 个属性的取值,x_{i2} 表示第 i 个样本第 2 个属性的取值。

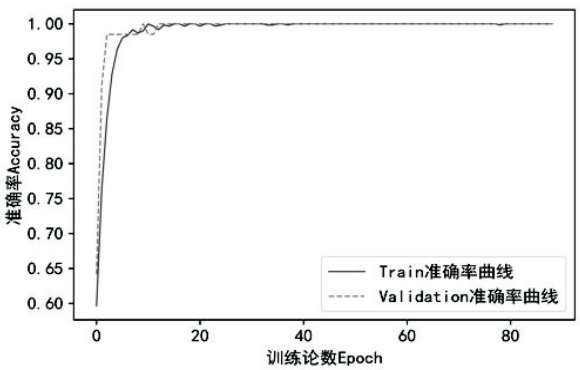
(2)数据构造。

陆空通话指令比较集以 $\{x_1, x_2, y\}$ 的形式构造,示例如下:x_1 和 x_2 为两个句子,y 为相似度标签,1 表示相似,0 表示不相似。比较集总计 27 452 对标注数据,训练集占 70%,验证集和测试集各占 20% 和 10%。其形式如下:

指令 1
指令 2
标签
示例:
CCA142,turn left heading 200 .
CCA142,turn left heading 200 for spacing.
1
CCA142,descend to 2100 meters.
CCA142,climb to 900 meters .
0

(3)模型训练。

随着训练轮数的增加,损失函数值呈下降趋势,准确率呈上升趋势,在训练轮数达到第 34 轮时,两者变化趋于平稳,准确率接近 99%,损失函数值降到了 0.002 8,模型基本达到收敛状态(见图 4)。



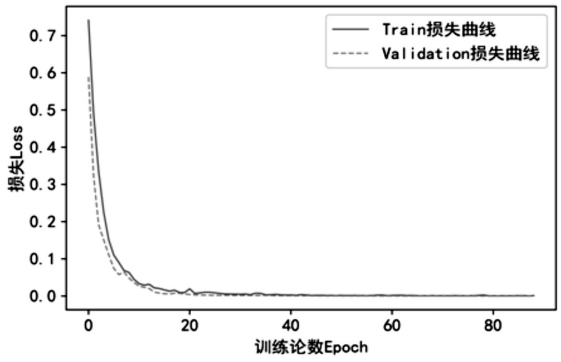


图 4 模型准确率、模型损失函数值的变化情况

(4)结果分析。

以相似度 0.9 作为判断相似与否的分界阈值,相似度高于 0.9 则认为两句话相似,反之不相似。三种方法在测试集上的准确率如表 3 所示。基于编辑距离的方法表现最差,由于陆空通话对飞行动作和动作数据有确切要求,因此在语言结构不变的基础上变换单词为反义词时语义正好相反,而该方法无法准确度量语义,造成相似度比较准确率低的结果。基于 wordnet 层级距离的方法虽能够识别同义词和反义词,但语言结构变化会引起相似度计算减小,造成判断准确率下降。以神经网络和词向量来计算句子相似度的方法取得了较好的收益,准确率达到了 93.6%,证明词向量是表征语义的良好手段,相对能更大限度蕴含词义和语言结构信息,作为网络输入能对下游管制员飞行员人机对话系统产生正面的影响。

表 3 句子相似度算法比较

方法	测试组数/正确组数/准确率
基于编辑距离	2 745/1 200/43.7%
基于 wordnet 层级距离	2 745/1 806/65.8%
基于 Siamese 网络	2 745/2 569/93.6%

4 结束语

以近阶段常规陆空通话为知识来源,将概念分类和句子相似度计算纳入词向量评价当中。概念分类的准确率平均值达 80.2%,浅层证明了词向量表征语义区分单词的特性。基于词向量的句子相似度计算准确达 93.6%,远超基于词义和语言结构的其他方法,进一步证实了词向量表征语义的功能,具备作为下游人机对话系统输入的条件。

研究存在两点不足:构造比较数据集时方法不够规范,耗时长覆盖少;人为设定阈值作为相似与否的分界存在较大主观性。后续工作研究重点将围绕数据集构造、相似分界阈值设定展开。

参考文献:

[1] 由 扬,徐肖豪. 空管模拟机的 IBM ViaVoice 技术实现研究[J]. 中国民航学院学报,2002,20(3):6-9.

[2] 刘万凤. 语音指令识别在陆空通话(英语)中的应用技术研究[D]. 南京:南京航空航天大学,2012.

[3] 卢薇冰. 基于 CNN 的陆空通话语义识别方法[D]. 天津:中国民航大学,2017.

[4] 路玉君. 基于 RNN 的陆空通话语义描述与度量方法[D]. 天津:中国民航大学,2017.

[5] MIKOLOV T, YIH W T, ZWEIG G. Linguistic regularities in continuous space word representations[C]//Conference of the North American chapter of the association for computational linguistics. Atlanta:[s. n.],2013:746-751.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//International conference on learning representations. Scottsdale, Arizona:[s. n.],2013:1-12.

[7] SCHNABEL T, LABUTOV I, MIMNO D, et al. Evaluation methods for unsupervised word embeddings[C]//Proceedings of EMNLP. Lisbon, Portugal: ACL Press, 2015: 298-307.

[8] ROGERS A, ANANTHAKRISHNA S H, RUMSHISKY A. What's in your embedding, and how it predicts task performance[C]//Proceedings of the 27th international conference on computational linguistics. New Mexico, USA:[s. n.], 2018:2690-2703.

[9] TULKENS S, EMMERY C, DAELEMANS W. Evaluating unsupervised dutch word embeddings as a linguistic resource [C]//Tenth international conference on language resources and evaluation. Portorož, Slovenia: European Language Resources Association,2016:4130-4136.

[10] SALTON G. A vector space model for automatic indexing [J]. Communications of the ACM,1975,18(11):613-620.

[11] HINTON G E, MCCELLAND J L, RUMELHART D E. Distributed representations[M]//Parallel distributed processing: exploration in the microstructure of cognition. Cambridge: MIT,1986:77-109.

[12] BARONI M, DINU G, KRUSZEWSKI G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors[C]//ACL 2014: the 52nd annual meeting of the association for computational linguistics. Baltimore, Maryland: ACL,2014:238-247.

[13] 刘 敏. 基于词向量的句子相似度计算及其在基于实例的机器翻译中的应用[D]. 北京:北京理工大学,2015.

[14] 陈丽莎. 自动问答系统中基于 WordNet 的句子相似度计算研究与实现[D]. 广东:华南理工大学,2015.

[15] MUELLER J, THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]//Proceedings of the thirtieth AAAI conference on artificial intelligence. Phoenix, Arizon: AAAI,2016:2786-2792.