广西科学院学报
Journal of Guangxi Academy of Sciences

2018,34(1):68～71
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-03-20 DOI:10.13657/j.cnki.gxkxyxb.20180320.006
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180320.1021.012.html

蒙古语词向量评测研究*
Research on Mongolian Word Vectors Evaluation

乌云塔那,王斯日古楞**
Wuyuntana, Wangsiriguleng

(内蒙古师范大学计算机与信息工程学院,内蒙古呼和浩特 010022)
(Computer and Information Engineering College, Inner Mongolia Normal University, Hohhot, Inner Mongolia, 010022, China)

摘要:词向量具有良好的语义特性,可用于改善和简化许多自然语言信息处理应用。本研究利用 CBOW 和 Skip-gram 两种模型架构在不同数据和不同维度下训练蒙古语词向量,然后结合蒙古语特征设计一个语义语法综合测试集,并在此测试集上用语义和语法相似度来评测词向量质量。研究结果表明,蒙古语语义和语法相似性任务上,Skip-gram 模型优于 CBOW 模型,Skip-gram 模型的窗口大小为 5 的情况下,词向量质量最好,且随着词向量维度或训练数据的增大,词向量质量有明显的提高。

关键词:词向量 CBOW 模型 Skip-gram 模型 词向量质量 语义语法相似度
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1002-7378(2018)01-0068-04

Abstract: The words vector has good semantic properties and can be used to improve and simplify many natural language processing applications. This study used CBOW (continuous Bag of words) and Skip-gram two model architectures to train the Mongolian word vectors in different data and different dimensions. Then we design a comprehensive semantic syntactic test set based on the Mongolia language features. And on this test set, we use semantic and syntactic similarity to estimate the quality of the word vectors. The results indicate that Skip-gram model is superior to CBOW model in Mongolian semantic and syntactic similarity tasks, and the word vectors quality is the best when the window size is 5. Moreover, with the increase of the word vectors dimension or training data, the quality of the word vectors is obviously improved.

Key words: word vectors, CBOW model, Skip-gram model, quality of the word vectors, semantic syntactic similarity

0 引言

要将自然语言交给机器学习中的算法来处理,通常需要将语言数学化,词向量就是将语言中的词进行数学化的一种方式。词向量将某种语言中的每一个词映射成一个固定长度的短向量,将所有这些向量放在一起形成一个词向量空间,而每一词向量则为该空间中的一个点,在这个空间上引入"距离",则可以根据词之间的距离来判断它们之间的语义、语法上的相似性。词向量具有良好的语义特性,可

收稿日期:2017-11-01
修回日期:2017-12-15

作者简介:乌云塔那(1992－),女,硕士研究生,主要从事蒙古文信息处理研究, E-mail:724831848@qq.com。

*内蒙古自治区自然科学基金项目"基于条件随机场的蒙古文命名体识别研究"(2016MS0623)和国家自然科学基金项目"基于神经网络的蒙汉机器翻译研究"(61762072)资助。

**通信作者:王斯日古楞(1970－),女,博士,教授,主要从事蒙古文信息处理研究,E-mail:siriguleng@imnu.edu.cn。

用于改善和简化许多自然语言信息处理应用,并且词向量的质量影响自然语言信息处理应用的性能。因此,针对蒙古语词性标注、命名实体识别、短语识别、机器翻译方面的应用需求训练蒙古语词向量,研究词向量的评测具有重要的研究意义。词语作为连续向量的表示具有悠久的历史。已有许多学者用不同模型训练了词向量,比较经典的模型有神经网络语言模型(NNLM)[1]、双对数线性语言模型(LBL)[2]、循环神经网络语言模型(RNNLM)[3]、连续词袋模型(CBOW)和 Skip-gram 模型[4]等。另外,Mikolov 等[5]还提出了 Skip-gram 模型的几个扩展,即 Hierarchical Softmax 算法、负采样算法和欠采样技术,从而提高了词向量的质量和训练速度。针对形态丰富的语言,Bojanowski 等[6]提出了一种基于 Skip-gram 模型的新方法。此外,词向量可用于改善和简化许多 NLP 应用[7-8]。

词向量的评价方法有两种:一个是把词向量融入到现有系统中,看能否提升现有系统[9];另一个是从语言学的角度分析词向量,比如相似度。研究者发现相似的词不仅彼此接近,而且这个词具有多重相似度。这在早期的变形语言中已被观察到,例如,英语名词可以有多个单词结尾,如果在原始向量空间的子空间中搜索类似的单词,可以找到具有相似结尾的单词[10-11]。蒙古语是黏着性语言,其形态丰富,并存在许多词根相同的名词和动词的变形形式,这些形式表示相似的概念。比如动词"ᠶᠠᠪᠤᠮᠤ, ᠶᠠᠪᠤᠭᠰᠠᠨ, ᠶᠠᠪᠤᠵᠤ, ᠶᠠᠪᠤᠭᠰᠠᠨ"都有相同的词根"ᠶᠠᠪ",表示"走"的不同形态。因此本研究从语言学角度的语义语法相似性来评价蒙古文词向量质量。蒙古语词向量评价方面不像英语一样有自己的语义语法测试集,因此,本研究结合蒙古语特征设计一个语义语法综合测试集,然后在此测试集上评测蒙古语词向量的质量。

1 蒙古语语义语法测试集的建立

词向量具有良好的语义特性,可以通过加减法操作来对应某种语义语法关系,并通过语义语法相似性来评价词向量。比如判断与 C(ᠡᠴᠡᠭᠡ) −C(ᠡᠵᠡᠨ) +C(ᠡᠮᠡᠭᠡᠳᠡᠢ)接近的向量是不是 C(ᠡᠮᠡᠭᠡᠳᠡᠢ)。

1.1 蒙古语语义测试集的建立

在大量数据上训练高维词向量时,所得到的向量可以回答诸如城市和它所属的国家之间的单词之间的微妙语义关系,例如,巴黎是法国的,柏林是德国的。根据这种现象本研究建立了两种蒙古语语义关系集:首都-国家关系和男-女关系。每个类别的两个实例如表 1 所示,相同类别的两个单词对连在一起构成一个语义问题,共有 100 个语义问题。本测试集中只包含一个词构成的单词,不包含多单词实体(如ᠨᠢᠭᠤ ᠮᠣᠰᠺᠤ)。

表 1 语义测试集中的两种语义关系实例

Table 1 Examples of two types of semantic questions in Semantic test set

Type of relationship	Word pair 1	Word pair 2
Capital city-Country	ᠪᠡᠶᠵᠢᠩ ᠬᠢᠲᠠᠳ	ᠰᠣᠸᠢᠸᠢᠶᠣ ᠣᠷᠤᠰ
Man-Woman	ᠡᠭᠴᠡ ᠡᠵᠡᠨ	ᠡᠬᠡ ᠡᠵᠡᠨ

1.2 蒙古语语法测试集的建立

蒙古语名词有格、数、领属等范畴的形态变化。蒙古语的格是通过名词后面缀接格附加成分来表示,例如"ᠭᠡᠷ ᠤ"(房子的),其中"ᠭᠡᠷ"(房子)是名词,"ᠤ" 是格附加成分,共有 7 种格附加成分。蒙古语的复数是在名词单数形式上缀接复数附加成分来表示。例如名词单数形式"ᠪᠠᠭᠰᠢ"(老师)上缀接复数附加成分"ᠨᠠᠷ",表示"老师们"的意思。蒙古语数词的变化也是在基数词上缀接各种附加成分表示,比如基数词"ᠨᠢᠭᠡ"(一)上缀接"ᠳᠤᠭᠠᠷ"附加成分形成"ᠨᠢᠭᠡᠳᠦᠭᠡᠷ",表示"第一"的意思。

本研究结合蒙古语法特征建立了关于蒙古语名词格、复数、数词和代词的 4 种语法关系集。每个类别的两个实例如表 2 所示,相同类别的两个单词对连在一起构成一个语法问题,共有 544 个语法问题。

表 2 语法关系测试集中 4 种类型语法问题实例

Table 2 Examples of four types of syntactic questions in Syntactic test set

Type of relationship	Word pair 1	Word pair 2
Numerals form change	ᠨᠢᠭᠡ ᠨᠢᠭᠡᠳᠦᠭᠡᠷ	ᠭᠤᠷᠪᠠ ᠭᠤᠷᠪᠠᠳᠤᠭᠠᠷ
Plural nouns	ᠪᠠᠭᠰᠢ ᠪᠠᠭᠰᠠᠨᠠᠷ	ᠣᠯᠠᠨ ᠣᠯᠠᠨᠠᠷ
Noun case	ᠰᠤᠮᠤᠨ ᠰᠤᠮᠤᠨ ᠳᠤ	ᠭᠡᠷ ᠭᠡᠷ ᠦ
Pronouns form change	ᠪᠢ ᠪᠢᠳᠡᠨ	ᠰᠢᠨᠡ ᠰᠢᠨᠡᠳᠡᠨ

1.3 词向量评测方法

使用类比方式在建立的语义语法测试集上评测蒙古语词向量质量。具体操作方法:设语义语法测试集中每个问题的 4 个词依次对应 a、b、c、d。已知 a 之于 b 犹如 c 之于 d。先给出 a、b、c,再看 C(a)−C(b)+C(c) 最接近的词是否是 C(d)。如果计算出来的向量与测试集中的词 d 完全相同,则认为是正

确答案。本研究评估所有问题类型的总体准确性，并分别对语义、语法问题进行评估。

2 蒙古语词向量模型架构

词向量可以使用 NNLM、RNNLM、LBL 等不同模型架构来学习，这些架构的计算复杂度比 CBOW 和 Skip-gram 模型要高出许多，因此本研究采用 CBOW 和 Skip-gram 模型训练蒙古词向量，并采用 Hierarchical Softmax 加速策略。

在 CBOW 模型和 Skip-gram 模型中，目标词 w_t 是一个词串中间的词而不是最后一个词，其拥有的上下文为前后 m 个词，m 为模型窗口的大小。

CBOW 模型是根据上下文预测目标语言的概率来优化词向量的。蒙古语 n-gram 语言模型中 4 元模型的效果最好，因此，本研究在窗口为 3～6 的情况下训练蒙古词向量，训练语料为 27 万蒙古文句子，词向量维度为 300，并用建立的语义语法测试集评测词向量。从表 3 可以看出，窗口为 4 的情况，下词向量质量最好，所以用于训练蒙古语词向量的 CBOW 模型的窗口大小为 4，目标词的上下文为该词前后的 4 个词。模型架构如图 1a 所示。举个例子，假如给定一个词的前面 4 个词“ / / / / ”和后面 4 个词“ / / / / ”，可以预测出这个词最大可能为“ ”。由此可以看出，在蒙古语中，当窗口设为 4 时，词语才能表达出良好的语义。

表 3 不同窗口下 CBOW 模型的性能

Table 3 Performance of CBOW model under different window size

Window size	Semantic-Syntactic accuracy (%)
3	23.30
4	24.14
5	19.71
6	14.91

Skip-gram 模型根据当前词预测上下文的概率来优化词向量，同样本研究在窗口为 3～8 的情况下分别训练蒙古词向量，训练语料为 27 万蒙古文句子，词向量维度为 300，并用建立的语义语法测试集评测词向量。从表 4 可以看出，窗口为 5 的情况下，词向量质量最好。

所以用于训练蒙古语词向量的 Skip-gram 模型的窗口大小为 5，根据当前词预测词前后 5 个词。模型架构如图 1b 所示。

表 4 不同窗口下 Skip-gram 模型的性能

Table 4 Performance of Skip-gram model under different window size

Window size	Semantic-Syntactic accuracy (%)
3	33.60
4	31.60
5	34.27
6	32.48
7	29.95
8	28.88

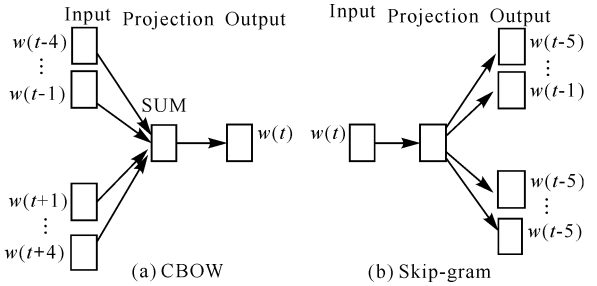

图 1 蒙古语词向量模型架构

Fig. 1 Mongolian word vectors model architecture

3 实例验证

使用上述 CBOW 模型和 Skip-gram 模型训练蒙古语词向量。训练语料采用 CWMT2015 的蒙古语训练语料和内蒙古大学 100 万词级的《现代蒙古文数据库》。上述两种语料是蒙古文拉丁转写形式，因此训练之前，先将单词与标点符号进行分割。

训练完成后，使用类比方式在建立的语义语法测试集上评测蒙古语词向量质量。表 5 是在不同大小的训练数据和不同维度下，用 Skip-gram 模型训练词向量的结果。由表 5 可以看出，在某些方面，添加维度或添加训练数据可以提高词向量质量。

表 5 Skip-gram 架构在蒙古语语义语法关系测试集上的总体准确性

Table 5 The overall accuracy of Skip-gram architecture in Mongolian semantic-syntactic relationship test set

Dimensionality	Accuracy (%)		
	12 M	17 M	23 M
100	17.91	22.88	25.36
200	20.57	25.54	28.11
300	22.35	26.53	29.04
600	21.22	27.54	28.26

为了比较两种模型架构，本研究在规模为 23 M 的训练语料上分别用 CBOW 模型和 Skip-gram 模型训练词向量，词向量维度为 300。结果表明，对于蒙古语而言，词法和语义任务上 Skip-gram 模型优

于 CBOW 模型(表 6)。

表 6 模型架构在语 X 语法测试集上的准确性比较

Table 6 The accuracy comparison of model architecture in semantic-syntactic test set

Model architecture	Semantic accuracy (%)	Syntactic accuracy (%)
CBOW	5.12	32.85
Skip-gram	15.12	42.95

综上所述,蒙古语语义准确率总体比较低,第一个原因是测试中忽略了同义词的概念,比如,就测试集中的"[illegible]"来说,"[illegible]" 有同义词"[illegible]",而本研究所采用的测评方法规定,只有计算出来的向量与测试集中的词"[illegible]"完全相同时才认为是正确答案,所以同义词"[illegible]"被忽略。第二个原因是训练语料库中缺少地名相关的词,因此训练出来的地名词向量质量比较差,无法表达出其语义。

4 结束语

本研究分别使用 CBOW 模型和 Skip-gram 模型训练蒙古语词向量,并在自己建立的语义和语法测试集上评测了词向量的质量。研究表明,利用 Skip-gram 模型且窗口为 5 的情况下蒙古语词向量质量最好。随着词向量维度或训练数据的增大,词向量质量有明显的提高。

蒙古语是个形态丰富的语言,有着丰富的数、格、时、体、态等形态变化。这导致了蒙古语的词汇量庞大。词向量训练时使用固定大小的词汇表,这使得罕见词语无法向量化。因此,在后续工作中将研究基于子词单元(比如词素级、字符级)的词向量表示来提升形态丰富语言的性能。词向量训练好以后,通常会作为各种神经网络结构的初始值,Word2vec 模型是很浅层的神经网络,词向量经预训练后做为其初始值,通常可以提升任务上的效果。因此,后续研究将会把训练好的词向量作为初始值,运用到蒙汉机器翻译任务上,以提升其翻译效果。

参考文献:

[1] BENGIO Y,DUCHARME R,VINCENT P,et al. A neural probabilistic language model[J]. Journal of Machine Learning Research,2003,3:1137-1155.

[2] MNIH A,HINTON G. Three new graphical models for statistical language modelling[C]. 24th Annual International Conference on Machine Learning (ICML),Corvallis,2007.

[3] MIKOLOV T. Statistical language models based on neural networks[D]. Lausanne: Brno University of Technology,2012.

[4] MIKOLOV T,YIH W,ZWEIG G. Linguistic regularities in continuous space word representations[C]. Proceedings of NAACL-HLT,2013.

[5] MIKOLOV T,SUTSKEVER I,CHEN K. Distributed representations of words and phrases and their compositionality [C]//BURGES C J C, BOTTOU L, WELLING M,et al(eds.). Advances in neural information processing systems 26 (NIPS 2013). Nevada: [S. n.],2013.

[6] BOJANOWSKI P,GRAVE E,JOULIN A,et al. Enriching word vectors with subword information[Z]. Facebook AI Research,2016.

[7] COLLOBERT R,WESTON J,BOTTOU L. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12:2493-2537.

[8] KIM Y. Convolutional neural networks for sentence classification[C]. Empirical Methods in Natural Language Processing. [S. l.],2014.

[9] TURIAN J,RATINOV L,BENGIO Y. Word representations:A simple and general method for semi-Supervised learning[C]. Proc Association for Computational Linguistics. [S. l.],2010.

[10] MIKOLOV T. Language modeling for speech recognition in Czech[D]. Lausanne: Brno University of Technology,2007.

[11] MIKOLOV T,KOPECKY,BURGET L,et al. Neural network based language models for highly inflective languages[C]. International Conference on Acoustics,Speech and Signal Processing. [S. l.],2009.

(责任编辑:陆 雁)