

文章编号: 1003-0077(2020)10-0033-06

基于词向量的藏文语义相似词知识库构建

龙从军^{1,3}, 周毛克², 刘汇丹³

(1. 中国社会科学院 民族学与人类学研究所, 北京 100081;

2. 中国社会科学院大学(研究生院), 北京 102488;

3. 中国科学院 软件研究所, 北京 100083)

摘要: 词向量在自然语言处理研究的各个领域发挥着重要作用。该文从语言学角度出发, 讨论了词向量技术与语言学理论的关系; 根据词向量的特征, 提出利用藏文词向量构建语义相似词知识库。该文以哈尔滨工业大学的《词林》为基础, 通过汉藏双语词典对译, 在获取对译词的词向量的基础上, 计算对译词的词向量与原子词群平均词向量的差值, 利用不同的差值, 自动筛选出与原子词群语义相似度较小的词。该文分别以藏文的词和音节为单位计算词向量, 自动筛出不属于原子词群的词, 通过对自动筛选结果与人工筛选结果对比, 发现两者具有较高的一致性, 这说明词向量计算结果与人的语言直觉具有较高的一致性。总体来说, 该文所采用的方法有助于提高藏文语义相似词知识库构建效率。

关键词: 词向量; 藏文; 语义相似词

中图分类号: TP391

文献标识码: A

Construction of Knowledge Base of Semantic Similar Tibetan Words Based on Word Vectors

LONG Congjun^{1,3}, ZHOU Maoke², LIU Huidan³

(1. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081, China;

2. Graduate School, University of Chinese Academy of Social Sciences, Beijing 102488, China;

3. Institute of Software, Chinese Academy of Sciences, Beijing 100083, China)

Abstract: Word vectors play an important role in various fields of natural language processing. This paper tries to reveal the relationship between word vector technology and linguistic theory. Based on the features of word vectors, this paper proposes an approach to construct knowledge base of semantic similar Tibetan words. Based on the Chinese <Cilin> dictionary, published by Harbin University of Technology, we compute the differences between every word vector and the average word vectors of the atomic word group. With the help of Chinese-Tibetan bilingual dictionary, we deploy such differences to select the similar words from word vectors by Tibetan words and Tibetan syllables, respectively. Compared with those of manual verification, we find that the results of word vector computing are consistent with human language intuition. This approach may improve the efficiency of constructing Tibetan knowledge base of semantic similar words.

Keywords: word vector; Tibetan; semantic similar word

0 引言

近年来,词向量(word vectors)成为自然语言处理研究领域备受关注的热点。词向量到底是什么?

单从字面上分析,就可以看出其是语言学和数学结合的产物。词是语言学中最基本的概念之一,是“最小的能够独立运用的语言单位”。向量则是数学中的基本概念之一(起源于物理学),是“具有大小和方向的量”。词和向量结合形成的“词向量”在一定程

收稿日期: 2019-09-16 定稿日期: 2019-10-26

基金项目: 中国社会科学院创新工程项目(2019MZSCX005);喜马拉雅区域协同创新中心项目(ZFYJY201901009)

度上能够表达一个特定的“词”在大小(文本域)和方向(上下文)上的向量总和,其中最典型的是词的部分分布信息和语义信息,可以通过特定的计算获得。要想了解计算出的词向量到底包含什么样的信息,有必要了解词的分布理论以及词义与词的分布之间的关系。

1 词的分布与词的意义

索绪尔把语言看成一个符号系统,这个符号系统由音位、音节、语素、词、短语、句子等要素按照一定的层级组成。他多次借用棋盘上的棋子来说明,进入语言符号系统中的各要素的价值是由系统组成成员各要素之间的相互位置来决定的,就像棋盘中的各棋子的价值由它们在棋盘上的位置决定一样^[1]。在索绪尔看来,语言符号系统各成员之间的位置关系可以概括为两种最基本的关系:组合关系和聚合关系。组合关系指两个同一性质的结构单位按照线性顺序组合起来的关系,组合关系是一种共现关系。聚合关系指在语言符号组合关系的某一位置上能够互相替换且功能相同和相似的一类符号之间的关系,聚合关系是一种替换关系。组合关系和聚合关系体现在语言符号系统的各个层级。美国结构主义语言学家布龙菲尔德提出“位置”和“形类”等概念,用来描写语言的分布特征。他在分析音位时采用了位置、替换、顺序、组合等术语,充分体现了分布思想^[2]。斯瓦迪士明确提出分布概念^[3];哈里斯界定了分布概念,并提出语言的分布结构,即“一个元素的分布将被理解为其所有环境的总和”;“首先,语言的各个部分彼此之间的关系不是任意发生的,在某种位置上,每个元素的发生都关联着某种其他的元素”;“如果我们认为词或词素 A 和 B 在意义上与 A 和 C 不同,那么我们会经常发现 A 和 B 的分布比 A 和 C 的分布更为不同”^[4]。这种分布假设不断被阐释。“在意义上相似的词出现在相似的上下文中”^[5];“如果有足够多的文本材料,相似意义的词将出现在相似的环境中”^[6];“一种能够足够好地捕捉到词在自然文本中如何使用的表征方式将可以捕获到足够多的它们所蕴含的意义”^[7];“发生在同样上下文语境中的词倾向有相同的意义”^[8]。

由此可见,意义的差异与分布的差异有关,换句话说,不同的意义关联着不同的分布^[9]。如果 A 和 B 有几乎相同的环境,我们说它们是同义词。如果 A 和 B 有一些共同的环境,也有一些不同的环境,

我们就说它们有不同的含义。意义差异的数量大致相当于它们环境的差异量。如果 A 和 B 从来没有相同的环境,我们就说它们是两个不同语法类的成员^[4]。这里需要强调的是利用词语的分布构造的模型,不是指词在人们头脑中所体现的那种概念模型,而是词在文本中跟位置(分布)距离相关的抽象模型,计算机与人对词的词义感知不同,但是计算出的词向量却与人对词的直觉类似。

2 利用词向量构建藏文语义相似词知识库

词向量模型可以把语义相似度高的词分组为集合,理论上可以自动构建语义相似词知识库。然而要获取一个性能较高的词向量模型需要大规模的语料和相对精确的分词工具,这两个前提在藏文及类似的低资源民族语言中成为一种奢望。藏文语料规模有限,分词精度不高,完全自动构建语义相似词知识库存在一定的困难。因此,我们设想在中文《词林》汉藏对译的基础上,采用词向量模型辅助构建藏文语义相似词知识库。

2.1 中文《词林》

本文所用的《词林》是由哈尔滨工业大学社会计算与信息检索研究中心的研究人员在梅家驹的纸质版《同义词词林》^[10]基础上删减扩充形成的电子版。电子版《词林》^[11]共包含 77 343 个词、90 102 个义项,把义项分成 12 个大类、97 个中类、1 400 个小类、4 026 个词群和 17 797 个原子词群。组织结构如下,“大类”用一个大写英文字母表示,“中类”在“大类”编码后添加一个小写英文字母,在“中类”编码后添加两位十进制整数构成“小类”编码,在“小类”编码之后附加一位大写英文字母构成“词群层”,在“词群层”编码之后附加两位十进制整数构成“原子词群编码”,基本结构如图 1 所示。

2.2 藏文《词林》

汉藏对照版《词林》雏形是利用双语词典根据中文《词林》中的词条匹配获得,存在如下一些问题。

(1) 中文词条无藏文对应。例如,“Aa01A02=人类 生人 全人类”该原子词群有三个词“人类”“生人”“全人类”,对应的“TAa01A02=མི་རིགས་ | [] | མི་རིགས་རྒྱུ་པ་”中只有两个词,中间的“生人”无对应的藏文词条。(TAa01A02 表示藏文《词林》的语义编码)

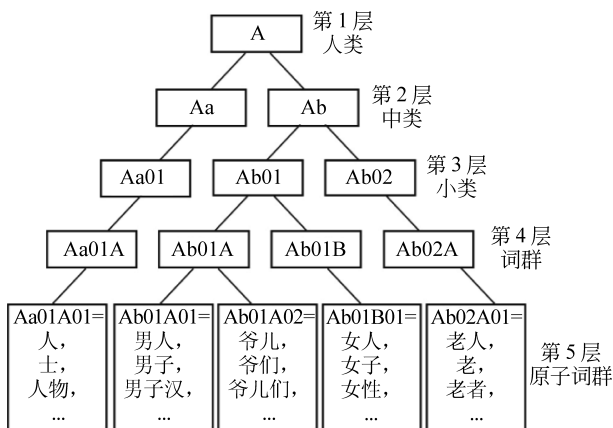


图 1 哈尔滨工业大学中文《词林》结构示意图

(2) 中文词条有多条藏文对应。例如, “Aa01A03=人手 人员 人口 人丁 口 食指”该原子词群有六个词, 对应的“TAa01A03 = ལས་མི/ ལས་ཀ/ ཕྱེད་མཉམ་ན་/ མི་ལྔ་/ མི་གྲངས་/ མི་འབྲེང་/ མི་འབྲེང་/ མི་གྲངས་/ མི་དང་མ་/ | [] || [] |”中有八个词, 其中, “人手”对应“ལས་མི/ ལས་ཀ/ ཕྱེད་མཉམ་ན་/”两个词, “人口”对应“མི་གྲངས་/ མི་འབྲེང་/”两个词, “人丁”对应三个词“མི་འབྲེང་/ མི་གྲངས་/ མི་དང་མ་/”, 而“口”和“食指”无藏文词条对应。

(3) 中文词条对应的藏文不恰当。例如, “Ac03C05=结巴 结子 咬舌儿 大舌头”对应的藏文词条为“TAc03C05 = ཁ་དྲིག་པ་/ཁ་དྲིག་དྲིག་པ་/ཐྱེ་དྲིག་དཔ་དྲིབ་སྒྲ་བ་/མཐུད་པ་/ཐྱག་མཐུད་/|ཐྱེ་སྐད་བྱག་པ་/ཐྱེ་དྲིག་སྒྲ་སྐད་མེ་བདེ་བ་/|ཐྱེ་ཐྱིག་ཐྱེ་དྲིག་/|”。该原子词群中的“结子”词条对应的藏文词条“མཐུད་པ་/ཐྱག་མཐུད་/”不恰当,这里“结子”采用了隐喻表达手法,整个词群的意义是指“说话表达不清,结巴”。“结子”词条对应的藏文是指“绳子等打结而形成的接头”。

汉藏文在表达现实世界和观念世界的概念时,存在一定的差异。这种差异是多层次的,从形式上看,音节之间可以形成多对多的关系。在汉语中,一个音节词可以对应藏文的多个音节词,反之亦然;从概念表达上看,汉语中的一个概念可以用藏文的多个词形来表达,反之亦然;从概念的所指来看,汉藏语言表达也存在差异,如“肉”在汉族地区通常指“大肉”,而在藏族地区通常指“牛羊肉”;还有词义在引申、转喻、褒贬色彩等方面存在差异,如“人丁”本意指“成年男子”,后泛指“人口”,对应的藏文词“མི་དང་མ”只表达“成年男子”之意,不泛指“人口”。

针对上述基于词条匹配的对译藏文《词林》出现的问题,我们采用词向量方法解决两个核心问题:

①删除多对多词条中语义偏差较大的藏文词条,如

问题(2)中的藏文词条“མི་དང་མ”；②删除原子词群中语义偏差较大的藏文词条，如问题(3)中的藏文词条“མདུན་པུ་ཕྱག་མདུན་”。

3 藏文词向量训练

2003 年 Bengio 等提出利用神经概率语言模型训练词的分布式表示,即词向量^[12]。词向量表示了语言的深层语义,用稠密的向量解决了传统 one-hot 表示带来的维度灾难和词汇鸿沟问题。2013 年 Mikolov 等改进了神经网络语言模型,提出包含 CBOW (continuous bag-of-words) 和 Skip-gram (continuous skip-gram) 两种计算框架的语言模型 Word2Vec,可实现在海量数据集上进行高效词向量训练^[13]。CBOW 原理是利用词前后 n 个词预测当前词;而 Skip-gram 则是利用当前词预测前后 n 个词。

英文词与词之间有空格,以英文为研究对象的词向量技术主要以词为单位训练向量,故称之为词向量。但是,对于汉、藏文来说,词与词之间没有明显边界标记,训练词向量之前需要对文本进行分词预处理,分词的准确性将影响词向量训练的结果。汉、藏文有音节边界,可以尝试以音节(音节和字之间大体存在一对一的关系)为单位训练音节向量(字向量)。本文采用了词和音节两种单位训练藏文的词和音节向量。

3.1 语料来源

藏文词向量训练使用了我们自己构建的互联网藏语篇章文本语料库,该语料库中的语料由从互联网上抓取的藏文文本构成,目前该语料库共包含44.09万篇藏文文本,共计949.88万句、2.28亿音节字。我们精选了其中规模比较大、文本质量高的13个网站的文本来训练词向量。

词向量训练时,使用卓玛拉藏文词法分析软件^①来对全部文本进行词语和音节切分,切分音节的时候将所有疑似黏着(紧缩)形式的音节都切开。文本材料详细来源如表1所示。

3.2 参数设置

本文实验中,我们使用相同的配置训练词向量和音节向量时。训练采用 CBOW 模型,维度设置为

① <http://tibetan.iea.cass.cn:8081/>.

表 1 训练词向量的文本情况

文本来源	文本总量/ MB
blog.amdotibet.cn	114.86
epaper.chinatibetnews.com	324.88
tb.chinatibetnews.com	544.80
tb.tibet.cn	136.48
tb.xzxw.com	119.99
ti.kbcmw.com	67.60
ti.tibet3.com	237.55
tibet.cpc.people.com.cn	27.20
tibet.people.com.cn	136.00
tibetan.qh.gov.cn	232.61
www.qhtb.cn	202.87
www.tibet.cn	114.80
xizang.news.cn	57.73
合计	2 317.37

50,上下文窗口大小设置为 5。最终训练得到的词向量文件共包含 106 085 个词语,音节向量文件共包含 14 912 个音节。

4 藏文《词林》自动构建实验及结果

4.1 基于词的词向量实验及结果

藏文《词林》雏形是基于汉藏双语词典匹配的结果,一部分词找不到对应词条,无藏文对译的词有 32 112 条,一部分词有多条藏文词对应,总共对译的藏文词条有 116 622 条。另外,藏文词与词之间无边界标记,一部分词切分不准确,影响词向量的结果;同时分词工具的词切分规则与词典中的收词原则不一致,一部分词可以通过对照词典对译,但在训练词向量的文本中没有出现,未能获得词向量结果,未获得词向量数据的词条有 74 657 条。获得词向量数据的词条有 41 965 条。

(1) 实验数据过滤:删除没有藏文对译的词条,例如,

Aa01A04= 劳力 劳动力 工作者

TAa01A04 = (ཤེད་གྲུགས།, 1) (ལལ་ལྷན་རྒྱུ་གྲུགས།, 0)
(ལལ་ལྷན་རྒྱུ་གྲུགས།, 1) | (, 0,) | (, 0,) |

“1”表示获得了词向量数据(Value),(, 0)表

示有对译词条,但未获得词向量数据,(, 0,)表示没有词条对应,无词向量数据。

例子中“劳动力”和“工作者”无对应藏文,删除。

(2) 词向量数据处理:①考察一(汉文词条)对多(藏文词条)的情况,获取多个藏文对译词条中语义相似度值,求出多条藏文词条词向量的中心点,并求出各词条词向量的值到中心点的余弦相似度值(WordValue1)。

②考察原子词群的词向量,求出整个词群的词向量的中心点,求出各词条到原子词群的词向量中心点的余弦相似度值(WordValue2);求出原子词群各词条词向量的平均值(WordValue3);原子词群中各词条的相似度值与平均值的差(WordValue2—WordValue3),最终获得的数据如表 2 所示。

表 2 基于词的藏文词原子词群向量示例

	词条	Word Value1	Word Value2	Word Value3	WordValue2— WordValue3
人手	མི་ལག	0.590	0.271	0.344	-0.074
	ལས་མི།	0.590	0.384	0.344	0.040
人员	མི།	0.617	0.318	0.344	-0.026
	མི་རྒྱ།	0.617	0.285	0.344	-0.060
人口	མི་ གཙང་ལ།	0.888	0.486	0.344	0.142
	མི་ འཛོལ།	0.888	0.400	0.344	0.056
食指	མཛུབ་ མོ།	0.595	0.111	0.344	-0.234
	ཁྱིམ་ མི།	0.595	0.303	0.344	-0.041

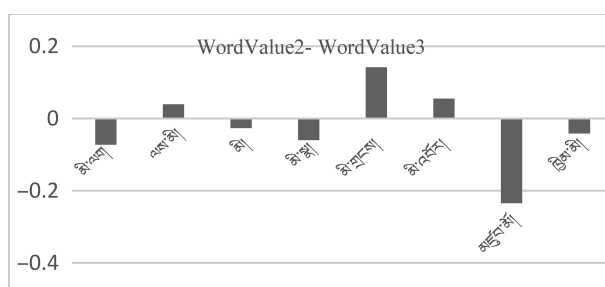


图 2 基于词的 WordValue2—WordValue3 示意图

表 2 中, WordValue2—WordValue3 值可以分成两个大类:正值和负值,正值则表示相似性较高,值越大,相似性越大;负值则表示相似性较小,值越小,相似性越小。从图 2 可以看出,མཛུབ་མོ的 WordValue2—WordValue3 差值较大,说明这个词与该原子词群的相似性较小,从人的语言直觉也能体会到这个词与原子词群中其他词的语义差距较大。

“食指”本义并不指“人口”，在特定语境中可以引申为“人口”，按照字面翻译成藏文后为“མཚུངས་མོ”，在藏文中不一定指“人口”的意思，由此可见，词向量计算的结果与人的直觉类似。

但是 WordValue2－WordValue3 的值为多少可以作为确定原子词群相似性大小的边界呢？通过观察得知，如果 WordValue2－WordValue3 值小于等于－0.05 时，可以作为基础标准，随着差值变小，排除的词条也减少，原子词群中各词条语义相似度限制变松，具体情况如表 3 所示。从表 3 数据可以发现，基于词的词向量计算结果，总词条和剩余词条数量较小。

表 3 差值取值范围与词条数量

总词条	筛出词条	差值(小于等于)	剩余词条
41 965	9 270	－0.05	32 695
	8 410	－0.06	33 555
	7 608	－0.07	34 357
	6 814	－0.08	35 151
	6 074	－0.09	35 891
	5 482	－0.10	36 483
	3 278	－0.15	38 687
	1 876	－0.20	40 089

4.2 基于音节的词向量实验及结果

以词为单位的词向量训练面临两个问题：藏文文本材料书写的规范性和藏文分词的准确性。用来训练藏文词向量的文本主要来自网络，书写存在许多问题；更重要的是当前藏文分词错误率较高，直接影响词向量训练的结果。另外从 4.1 节可以发现，有 74 657 条藏文对译词条未获得词向量，说明对译的词条和从文本分词中获得的词条之间有较大的差距。藏文音节之间有标记，采用音节为单位训练音节向量可以避免分词错误导致的词向量训练不准确的问题。再由音节向量获得词向量，即可有效避免藏文对译词条无词向量问题。

4.1 节中谈到总共对译的藏文词条有 116 622 条，以音节为单位训练音节向量，音节向量相加求平均值获得音节所在词的词向量，基于音节的藏文词向量原子词群示例如表 4 所示，示意图如图 3 所示。

如表 4 中 WordValue2－WordValue3 值同样分成两大类，正值和负值，正值表示该词与原子词群

中其他词相似性较高，值越大，相似性越大；负值则表示相似性较低，值越小，相似性越小。

表 4 基于音节的藏文词向量原子词群示例

	词条	Word Value1	Word Value2	Word Value3	WordValue2－WordValue3
人手	མི་ལག	0.280	0.237	0.299	－0.062
	ལས་མི།	0.403	0.367	0.299	0.068
	ལས་ཀྱི་བྱིད་མཁམ།	0.280	0.209	0.299	－0.090
人员	མི།	0.636	0.496	0.299	0.197
	མི་མུ།	0.578	0.425	0.299	0.126
	ལས་ཀྱི་བྱིད་མཁམ།	0.280	0.209	0.299	0.090
人口	མི་གཙམ།	0.589	0.349	0.299	0.050
	མི་འཛོལ།	0.794	0.496	0.299	0.197
人丁	མི་འཛོལ།	0.611	0.496	0.299	0.197
	མི་གཙམ།	0.449	0.349	0.299	0.050
	མི་དར་མ།	0.223	0.169	0.299	－0.130
口	རང་ཁ་བྱིད་ད།	0.296	0.180	0.299	－0.119
	རང་ཁ་གསོ་བ།	0.274	0.171	0.299	－0.128
食指	མཚུངས་མོ།	0.150	－0.025	0.299	－0.324
	བྱི་མ་མི།	0.288	0.3598	0.2998	0.060

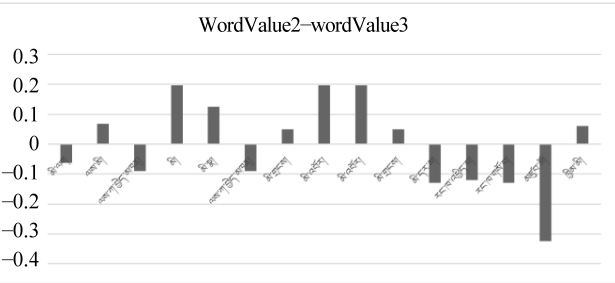


图 3 基于音节的 WordValue2－WordValue3 示意图

从图 3 可以看出，མཚུངས་མོ的 WordValue2－WordValue3 差值同样较大，说明这个词与该原子词群的相似性较小，这个结果与基于词的词向量是一致的。同样，以取值为多少作为确定删除词条的标准，也要根据具体情况来确定。表 5 分别列举了一些取值范围及词条删除情况。从表中可以看到，基于音节的词向量计算得到的总词条和剩余词条比较多。

4.3 基于词和音节词向量的语义相似词知识库构建实验比较

从 4.1 和 4.2 节可以看出，采用不同单位训练词

表 5 差值取值范围与词条数量

总词条	筛出词条	差值(小于等于)	剩余词条
116 622	20 262	-0.05	96 360
	16 323	-0.06	100 299
	13 097	-0.07	103 525
	10 474	-0.08	106 148
	8 335	-0.09	108 287
	6 677	-0.10	109 945
	3 278	-0.15	113 344
	770	-0.20	115 852

向量,其构建的知识库的词条总量有较大的差别,以词为单位的方法获得 41 965 词条,经过筛选之后,词条剩余最多 40 089 条,最少 32 695 条。以音节单位的方法获得 116 622 词条,经过筛选之后,词条剩余最多 115 852 条,最少 96 360 条。从数量上看,基于音节的词向量训练方法较好。为了考察自动筛选的可信度,本文把自动筛选结果与母语人语言直觉筛选结果进行了对比。以差值小于等于 -0.15 为标准,从两种不同方法获得的结果中自动筛选出语义相似度小的词条,让母语人做判断,自动筛选与语言直觉相符的画勾,不符合的画叉。筛选情况如表 6 所示。

表 6 自动筛选与人工筛选对比

	训练词向量的类型	原子词群数量	自动筛选词条数量	符合语言直觉词条数量	人机筛选一致率/%
Test1	词	775	320	179	55.93
	音节	778	139	90	64.74
Test2	词	428	194	132	68.04
	音节	810	146	100	68.49

从表 6 可以看出,自动筛选结果与人的语言直觉具有较高的相似性,以词为单位的词向量方法最高一致性达 68.04%,以音节为单位的词向量方法最高一致性达 68.49%。两个测试结果表明以音节为单位的词向量方法要比以词为单位的词向量方法好。同时也可以看出,不同人的语言直觉也存在一定的差距。

5 结语

词的分布与词的意义之间关系紧密,词的意义

可以通过其分布才能表现出来。自然语言处理面临着许许多多的歧义,既有结构上的,也有语义上的,要消除这些歧义只依靠词本身是难以解决的,还需要考察本词与他词之间的关系,获取更多词语之间的关系数据,以此来辅助解决歧义问题。基于语义分布假说理论的词向量计算技术正是获取了词与其共现词、词与其语义相似词之间的关系距离,以此为基础开展的自然语言处理研究获得了不错的效果。本文以哈工大中文《词林》为基础,开展藏文语义相似词知识库构建,通过自动筛选和人工筛选对比发现,词向量计算结果与人的直觉类似,充分说明可以利用词向量来构建藏文语义相似词知识库,在民族语言研究人力和财力不足的情况下,以自动构建和人工辅助相结合,加快知识库建设进度是当前和今后可行的研究方法。藏文文本材料较少,获取的词向量还存在不少问题,从构建的知识库来看,以词为单位的词向量获取的知识库词条少,与人的语言直觉一致性稍差;以音节为单位的词向量获取的知识库数量较大,与人的语言直觉一致性较高。基于音节的词向量训练方法,避免了分词的错误;同时,藏文音节独立成词的比例较高,基本上也是“名副其实”的词向量。但是也可以看到,自动筛选的结果有待进一步改善,除了扩大语料规模之外,还需要在词向量训练和原子词群之间各词条语义关系计算方面不断改进,同时也尝试把词向量技术应用到汉藏语言关系的亲疏研究中。

参考文献

- [1] 菲尔迪南·德·索绪尔,高名凯,译. 普通语言学教程[M].北京:商务印书馆,1999.
- [2] Harris Z S. Methods in structural linguistics [M]. Chicago & London: The University of Chicago Press, 1951.
- [3] Swadesh M. Thonemic principle [J]. Language, 1934 (10): 117-129.
- [4] Zellig S. Harris. Distributional Structure[J]. Word, 1954, 10(2-3):146-162.
- [5] Rubenstein H, Goodenough J. Contextual correlates of synonymy[J]. Communications of the ACM, 1965, 8 (10): 627-633.
- [6] Schütze H, Pedersen J. Information retrieval based on word senses [C]//Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. 1995: 161-175.

(下转第 50 页)

- 研究与应用[J]. 计算机技术与发展, 2008, 18(7): 233-236.
- [24] 中国社会科学院语言研究所词典室. 现代汉语词典, 第 6 版[M]. 北京: 商务印书馆, 2012.
- [25] 郭庆光. 传播学教程(第 2 版)[M]. 北京: 中国人民大学出版社, 2011.
- [26] Speer R, Havasi C, Lieberman H. Analogy Space: Reducing the Dimensionality of Common Sense Knowledge[C]//Proceedings of the 23rd AAAI Conference on Artificial Intelligence. 2008.
- [27] Tandon N, De Melo G, Weikum G. Acquiring comparative commonsense knowledge from the web[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014.
- [28] Boia M, Musat C C, Faltings B. Acquiring commonsense knowledge for sentiment analysis through human computation[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence. 2014.
- [29] Cambria E, Nguyen T V, Cheng B, et al. GEC-KA3D: A 3D Game Engine for Commonsense Knowledge Acquisition[C]//Proceedings of the 29th International Flairs Conference. 2016.



王亚(1988—), 博士, 工程师, 主要研究领域为常识获取、人工智能。

E-mail: wangya@ict.ac.cn



曹存根(1964—), 博士, 研究员, 主要研究领域为大规模知识处理、人工智能。

E-mail: cgcao@ict.ac.cn

(上接第 38 页)

- [7] Landauer T, Dumais S. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge[J]. Psychological Review, 1997: 104 (2), 211-240.
- [8] Pantel P. Inducing ontological co-occurrence vectors [C]//Proceedings of the 43rd Conference of the Association for Computational Linguistics, ACL'05 2005: 125-132. Morristown, NJ, USA: Association for Computational Linguistics.
- [9] Harris Z. Distributional Structure[C]//Proceedings of the papers in structural and transformational linguistics, 1970: 1775-794.
- [10] 梅家驹, 竺一鸣, 高蕴奇, 等. 同义词词林[M]. 上海: 上海辞书出版社, 1983.
- [11] 哈工大社会计算与信息检索研究中心. 同义词词林扩展版[EB/OL]. [2019-09-13]. <http://www.data-tang.com/data/42306/>
- [12] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3 (6) :1137-1155.
- [13] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [J/OL]. arXiv preprint arXiv: 1301.3781v3, 2013.



龙从军(1978—), 博士, 副研究员, 主要研究领域为藏语计算语言学。

E-mail: longcj@cass.org.cn



周毛克(1995—), 硕士研究生, 主要研究领域为藏语自然语言处理。

E-mail: zmk_cass@126.com



刘汇丹(1982—), 博士, 副研究员, 主要研究领域为多语言处理。

E-mail: huidan@iscas.ac.cn