一种基于相似度的藏文词同现网络构建及特征分析

加羊东周[1,3,4] **才智杰**[1,2,3,4] **才让卓玛**[1,2,3,4] **三毛措**[1,3,4]

1. 青海师范大学计算机学院，青海西宁 810016；

2. 西南民族大学计算机科学与技术学院，四川成都 610041；

3. 藏文信息处理教育部重点实验室，青海西宁 810008；

4. 青海省藏文信息处理与机器翻译重点实验室，青海西宁 810008

358521688@qq.com Czjqhsd@163.com cr-zhuoma@163.com 2627996852@qq.com

摘要

语言文字是人类智慧和文明的结晶，是经过漫长演化形成的复杂系统。语言同现网络采用复杂网络技术研究语言的特征，揭示语言文字的内部结构关系。文章通过分析相似性同现网络构建模块结构，提出一种基于相似度的藏文词同现网络构建方法，该方法以词为网络节点，以相似词间连边构造词同现网络。基于相似度藏文词同现网络构建方法，在大、中、小三类文档上建立了词同现网络，并分析了它们的统计特征，实验数据表明建立的藏文词同现网络都具有小世界效应和无标度特征。

关键词： 自然语言处理； 藏文； 词向量； 相似度； 同现网络

A Research on Construction and Feature Analysis of Similarity-based Tibetan Word Co-occurrence Networks

Jia Yang-dongzhou[1,3,4] **Cai Zhi-jie**[1,2,3,4] **Cai Rang-zhuoma**[1,2,3,4] **San Mao-cuo**[1,3,4]

1.College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016;

2.School of Computer Science and Technology,
Southwest Minzu University, Sichuan Chengdu 610041,China;

3.Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Qinghai Xining 810008;

4.Key Laboratory of Tibetan Information Processing,
Ministry of Education, Qinghai Xining 810008

358521688@qq.com Czjqhsd@163.com cr-zhuoma@163.com 2627996852@qq.com

Abstract

As the crystallization of human wisdom and civilization, Language is a complex system formed after a long evolution. Language concurrency network utilizes complex network techniques to study the linguistic features and reveal the internal structure of languages. In this work, we analyze the modular structure co-occurrence network and proposes a similarity-based method for constructing Tibetan word co-occurrence networks, in which Tibetan words serve as the nodes, and the similarity metrics among words serve

第十九届中国计算语言学大会论文集，第509页-第517页，海南，中国，2020年10月30日至11月1日。

as the edges. We established of similarity-based word co-occurrence network on three types of documents in terms of size, namely, large, medium and small. and analyzed their statistical features. The experimental data indicated that the Tibetan word co-occurrence networks have small-world effects and scale-free features.

Keywords: Natural Language Processing , Tibetan , Word Embedding , Similarity , Co-occurrence Network

1 引言

语言文字是人类智慧和文明的结晶，是经过漫长演化形成的复杂系 (Steels, 2000)。语言同现网络采用复杂网络技术定量考察和分析语言的特性, 验证语言同现网络的小世界效应和无标度特征, 揭示语言的内部结构关系。词同现网络是语言同现网络的一种表现形式 (孙文俊等, 2010), 揭示词与词之间的内部结构关系。词同现网络的定义不同，其构建词方法也各不相同。词同现网络构建方法主要有 n 阶 Markov 同现模型和相似性同现模型 (才智杰, 2018) 等两种。由于 n 阶 Markov 同现模型理论相对成熟且操作便捷，成为构建词同现网络的常用方法。

近年随着神经网络技术的飞速发展，词向量表示性能得到了显著提升 (才智杰, 2020)，方便了词相似度的计算，从而为构建相似性词同现网络奠定了理论基础。为了研究相似性同现网络模型的构建技术及验证相似性同现模型下藏文词同现网络的小世界效应和无标度特征，揭示藏文词同现网络的内部结构，本文在已有藏文词向量表示的基础上，研究了相似性模型的藏文词同现网络构建技术，提出了一种基于相似度的藏文词同现网络构建方法，并分析了它们的统计特征，实验数据表明建立的藏文词同现网络都具有小世界效应和无标度特征。

2 研究现状

自 Cancho 和 Solé (2001) 首次将复杂网络的方法引入语言研究中，学者们开始针对不同的语言从不同层面研究语言网络。基于 n 阶 Markov 同现模型依据句子中两个词的 n 阶 Markov 链建立语言同现网络，通过上下文的顺序制约关系揭示词间的关系；相似性同现模型通过词之间的相似性建立网络，通过节点的相似度和上下文语义关系揭示词间的关系。Barabasi (2002) 采用 n 阶 Markov 同现模型建立了英文词的同现网络，梁伟 (2012)、林枫 (2012) 和刘知远 (2007) 等采用 n 阶 Markov 同现模型建立了汉文字/词的同现网络。梁伟等 (Wei et al., 2012; Liang et al., 2015; Liang et al., 2009) 从文学的视角，通过词同现网络对中国、英、美作品做了一系列的比较研究工作；耿志杰 (2010) 等构建了图书情报领域关键词同现网络，并进行了结构研究；余传明 (2010) 等利用网站评论数据构建情感词汇同现网络，并挖掘情感词汇之间的关系及内部规律；Liu (2011) 从地理系统科学数据中提取关键词进行词同现网络分析和可视化；He (2016) 通过构建 6000 首华语流行歌曲歌名的词同现网络，揭示流行歌曲独特的词同现网络特征；李亚星 (2016) 等根据微博语料的特点，通过词同现网络获取关联性强和具有潜在传播效应的词语；Atsushi Tsuya (2014) 用词同现网络分析癌症病人日常发布与疾病相关的推荐信息获得到对病人需求的深层理解。以上文献都以 n 阶 Markov 同现模型构建词同现网络，迄今为止未见有关采

用相似性同现模型构建词同现网络的文献报道。刘知远 (2008) 等采用相似性同现模型建立了汉语依存句法网络，Cancho 和 Motter (Cancho and Sole, 2001; Cancho et al., 2004; Motter et al., 2002) 采用相似性同现模型分别建立了英语句法网络和概念网。

少数民族语言同现网络的研究相对较少，才智杰 (2018) 等采用 n 阶 Markov 同现模型在诗歌、散文、政治、佛教、教材、口语等不同类型的语料上构建了 97 个藏文字同现网络，分析了藏文字同现网络的最短路径长度、聚类系数和度分布，实验数据显示 97 个藏文字同现网络都具有小世界效应和无标度特性，表明藏文字同现网络都具有小世界效应和无标度特性。藏文词同现网络的研究未见文献报到。本文在已有藏文词向量表示的基础上，研究了相似性模型的藏文词同现网络构建技术，提出了一种基于相似度的藏文词同现网络构建方法，该方法以词为网络节点，以相似的词间连边构造词同现网络。在大、中、小三类文档上建立了基于相似度藏文词同现网络，并分析了它们的统计特征，实验数据表明建立的藏文词同现网络都具有小世界效应和无标度特征。

3 基于相似度的藏文词同现网络构建

3.1 基于相似度的藏文词同现网络构建模块结构

在语言学领域，词与词之间的关系具有很强的规律，词同现网络的文本表示可以捕捉文本结构信息，揭示其内在的组织原则与语言学规律。近年来随着深度神经网络的飞速发展，词向量在自然语言处理领域得到了广泛应用。词向量作为处理下游任务的输入特征，使下游任务的性能得到了显著改进和提升。

相似性同现模型是通过词之间的相似性建立网络，网络中的节点为词，同一文档中最相似两词对应的节点间连接一条边。即同现网络 $G = \{V, E\}$，$V = \{v_i | v_i \in T\}$，$E = \{e_i | e_i = \mathbf{max_j sim(v_i, v_j)}\}$，其中 T 表示藏文词的集合。基于相似度的藏文词同现网络构建结构如图 1 所示。

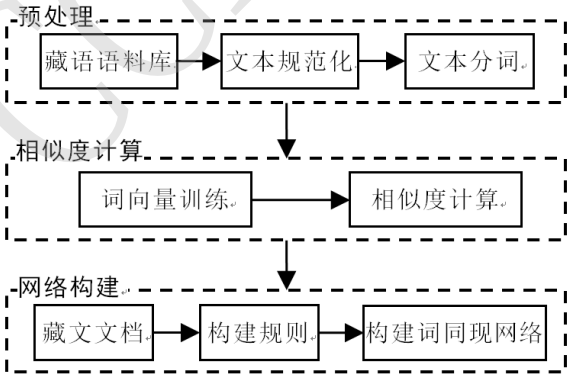

图 1: 基于相似度的藏文词同现网络构建结构

基于相似度的藏文词同现网络构建结构包含预处理模块、相似度计算及网络构建模块。预处理模块将语料库中的藏文本进行规范化和分词处理，通过规范化处理得到纯净的藏文文本数据，使用分词软件对藏文文本进行分词。相似度计算模块先训练词向量，将藏文词表示为向量，通过词向量计算出词之间的相似度。通常使用余弦相似度、欧氏距离计算词之间的相似度。词同现网络构建模块根据网络构建规则建立揭示词之间关系的网络，网络的节点为文档中的词，满足相似条件的两个词间连一条边。

3.2 基于相似度的藏文词同现网络构建规则

相似度计算模块和网络构建模块是基于相似度的藏文词同现网络构建的要素。词相似度的计算性能取决于词向量的效果，训练词向量的语料越大词向量效果越好，因此选用一个大语料训练词向量较合适 (我们用全集语料训练了词向量)。网络构建模块通过节点的选取和节点之间的连边规则构建网络，节点应该从当前文档中选取。连边规则决定词同现网络边的选择，揭示词之间的关系，是整个词同现网络构建的核心，称为词同现网络构建规则。

构建基于相似度词同现网络的规则中，需要考虑两个问题，其一是对于给定的节点词 A，如何选取该节点词的相邻节点词 X，其二是对于给定的节点词 A 已经选择了相邻节点词 B 的情况下，又选到相邻节点词 B 时该如何处理。问题一中相邻节点词 X 的选择可根据具体情况取与节点词 A 相似的前 n 个词。由于词的相似性具有传递性，当 $n=1$ 时具有 N 个节点词的词同现网络由 $\frac{N}{2}$ 个连通子图组成 (如图 2 所示)，不符合实际需求，因此在实际构建词同现网络时 $n \geq 2$ 较合适。

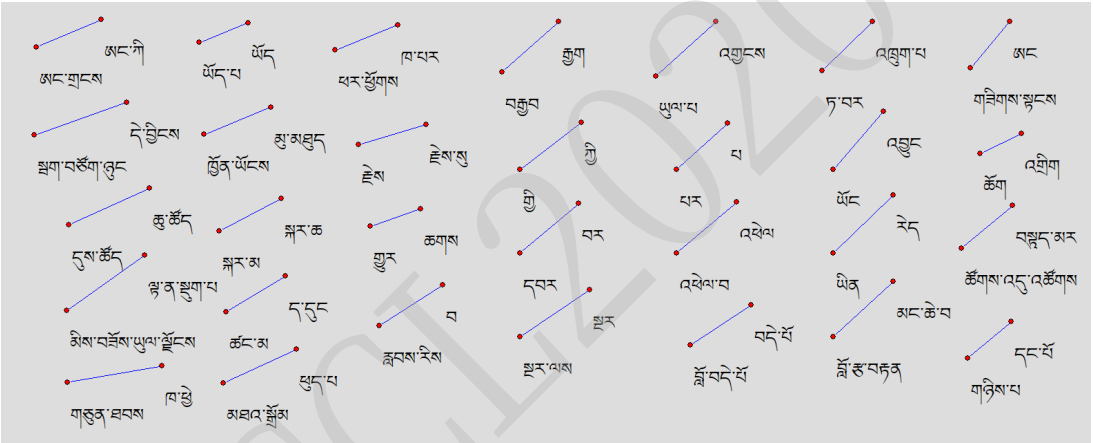

图 2: $n=1$ 时的词同现网络示意图

对于给定的节点词 A 已经选择了相邻节点词 B 的情况下，又选到相邻节点词 B 时有两种处理方法：第一种处理方法是忽略相邻节点词 B，节点词 A 和与它相邻的 $n-1$ 个相邻节点相连；第二种方法是忽略相邻节点词 B，增加与节点词 A 第 $n+1$ 个相似的词 C 为相邻节点，从而保持每个与节点 A 相邻的节点数为 n。在构建基于相似度词同现网络时选择第二种方法较合适。事实上，通过实验观察到采用第一种处理方法会使后继节点词的邻接节点越来越少，从而使词同现网络变为边稀疏网络，不能更好的反映词之间的内部结构关系。通过以上讨论可得以下基于相似度的藏文词同现网络构建规则和算法。

基于相似度的藏文词同现网络构建规则: 选用一种比较好的词向量训练方法和大的语料训练得到一个性能好的词向量表用于计算词相似度，给定文档中的每个词为网络节点 v_i，求出与节点 v_i 最相似度的 n 个节点 $u_j(j=1,2,...,n)$。若这 n 个节点 u_j 都与 v_i 不相邻，则连接 v_i，u_j，即 $(v_i\ u_j)$ 加入 E; 否则用节点 v_i 的第 n 个相似节点之后且与 v_i 未连接的点 u_k 替换节点 u_j。具体算法如下:

算法 1 基于相似度的藏文词同现网络构建算法

输入: 词向量矩阵 M^{N*D}，选定的藏文文本 T;

输出: 基于相似度的藏文词同现网络;

for $i = 1 \rightarrow |V|$ **do**
  for $j = 1 \rightarrow |V|$ **do**
    if $v_i == v_j$ **then**
      $similarity\_list[j-1] = 0$ v_i 和 v_j 为同一个词时, 相似度置为 0
    else
      $x \leftarrow B_{v_iM}$ 取 x 为 v_i 的词向量，B_{v_i} 为 v_i 的 one-hot 向量
      $y \leftarrow B_{v_jM}$ 取 y 为 v_j 的词向量，B_{v_j} 为 v_i 的 one-hot 向量
      $similarity\_list[j-1] \leftarrow similarity(x, y)$ 计算相似度并保存到数组中
    end if
    设置 n
    while $j < n$ **do**
      $index \leftarrow argmax(similarity\_list)$ 取出与 v_i 相似度最高值的下标
      $u_j \leftarrow V[index]$
      $similarity\_list[insex] = 0$
      if $(v_i, u_j) \in Eor(u_j, v_i) \in E$ **then**
        $n = n + 1$
        $j = j + 1$
      else
        $E \leftarrow (v_i, u_j)$
        $j = j + 1$
      end if
    end while
    $G \leftarrow (V, E)$
  end for
end for

3.3 基于相似度的藏文词同现网络构建

在构建藏文词同现网络时我们从青海师范大学建立的藏语语料库中选取了 18.07M 大小的语料，对其进行了预处理 (才智杰, 2018; 才智杰等, 2011)，将其作为词向量训练语料。词向量训练语料（下文称全集语料）信息见表 1。

	文学	政论	藏医	共计
大小	6.64M	8.53M	2.90M	18.07M
词条数	485815	542230	230935	1258980

表 1: 词向量训练语料信息表

由于藏文语料库相对较小, 因而我们参考文献 (才智杰等, 2019) 选用了在小规模语料库中表

现较好的 CBOW 模型训练词向量。CBOW 模型参数见表 2。

Dimsize	Window	alpha	Iter	hs	TWordSim215
300	5	0.025	500	0	47.67

表 2: CBOW 模型参数表

词向量的相似度通常由余弦相似度、欧氏距离表示，余弦相似度更能刻画向量间的相似程度。我们在计算相似度时采用词向量的夹角余弦值来评估词间的相似度，余弦相似度计算公式如下:

$$\mathbf{CosineSimilarity(u, v)} = \frac{\mathbf{u \cdot v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \tag{1}$$

其中 $u \cdot v$ 是两个向量的点积，$\|u\|_2$ 是向量 u 的范数。余弦相似度的取值在 $[0, 1]$，当余弦相似度的值越大表示两个向量越相似。

为了观察词同现网络的效果及特征分析，我们选用了大、中、小三个文档采用 2.2 节的词同现网络构建规则构建了藏文词同现网络，其中 n 取 2。大文档指用于训练词向量的大小为 18.07M 全集语料，中文档指从大文档中文学、政论、藏医等三类中各取选了 50% 而得到的大小为 8.54M 的语料，小文档指从大文档中任意抽取的大小为 2.08M 的语料。构建同现网络文档信息见表 3，构建的词同现网络如图 3 所示。

	大文档	中文档	小文档
大小 (M)	18.07	8.54	2.08
词条数	1258980	599407	165740

表 3: 构建词同现网络语料信息表

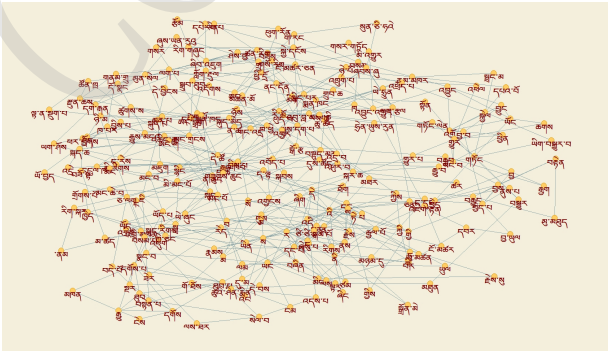

图 3: 藏文词同现网络示意图

4 藏文词同现网络的特征分析

4.1 小世界效应

为了从多方位考察藏文词同现网络的特征，我们对大、中、小三类文档 (文档信息见表 3，网络构建规则中 $n = 2$) 建立的基于相似度的藏文词同现网络利用 Pajek 网络分析工具进行了特征

统计分析, 同现网络基本数据对比表如表 4 所示，其中藏文字同现网络的统计参数来自文献 (才智杰, 2018)，汉文字同现网络的统计参数来自文献 (梁伟等, 2012)，汉文词同现网络的统计参数来自文献 (刘知远等, 2007)。

类型		N	E	D	$< k >$	L	Lr	$C(\%)$	$Cr(\%)$
藏文字 (均值)		3194	41943	7	26.2636	2.5644	2.4690	11.5398	0.8225
汉文字 (均值)		4520	96512	9	42.7000	2.4900	2.2400	38.0700	0.9500
汉文词		157000	8300000	–	64.35	2.63	2.99	0.619	0.00025
藏文词	大文档	38636	77272	15	4	7.4910	7.6204	0.2075	0.000048
	中文档	24047	48094	13	4	7.1841	7.2784	0.1691	0.000055
	小文档	12371	24742	13	4	6.8142	6.7987	0.1280	0.000017

表 4: 同现网络基本数据对比表

表中 N 表示词同现网络的顶点数、E 表示边数、D 表示直径、$< k >$ 表示平均度、L 表示平均最短路径长度、Lr 表示平均最短路径长度参照系数、C 表示平均聚类系数、Cr 表示平均聚类系数参照系数。

以上的实验数据体现了基于相似度的藏文词同现网络的以下特征:

(1) 在藏文词同现网络的统计参数中所有统计参数比较稳定，只是随语料大小的变化有小的波动，并不随语料大小的变化而有较大的变化，说明选取语料规模的大小对基于相似度的藏文词同现网络的统计参数没有太大的影响。

(2) 直径 D 的值在小语料集和大语料集上几乎相同，比汉文字/词、藏文字的大；藏文词的平均度 $< k >$ 都为 4，远远小于汉文字/词、藏文字的平均度。说明基于相似度的藏文词同现网络是一种边稀疏网络，相似词之间的关联度较弱。

(3)3 个藏文词同现网络都具有小的平均最短路径 L，且 $L \approx Lr$，$C \gg Cr$，说明基于相似度的藏文词同现网络具有小世界效应。

4.2 无标度特性

我们分析了构建的 3 个藏文词同现网络的度分布情况，与其他语言网络的度分布类似，网络中少数的节点往往拥有大量的连接，而大部分节点却很少，这些大多数点对无标度网络的运行起着主导作用，呈现“胖尾”现象。说明藏文词同现网络的度分布服从幂律分布，显示了无标度特性。双对数坐标下三类文档的词同现网络度的分布见图 4。

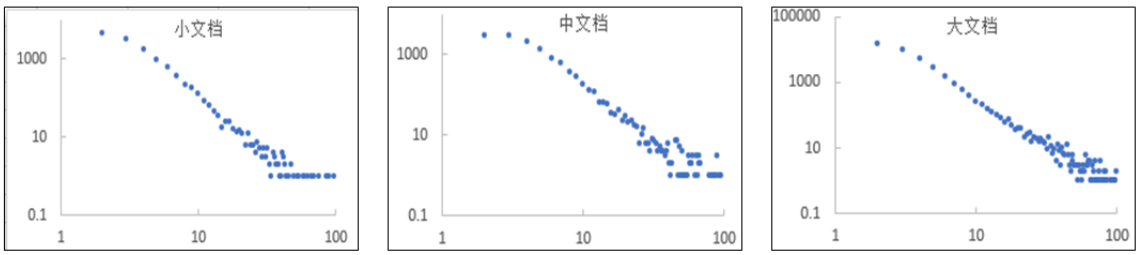

图 4: 双对数坐标下三类文档的度分布图

5 结论

语言同现网络通过复杂网络的方法研究语言网络的特征，有助于揭示语言文字的内部结构关系。词同现网络是语言同现网络的一种表现形式，其构建方法主要有 n 阶 Markov 同现模型和相似性同现模型等两种。学者们已从不同角度研究了基于 n 阶 Markov 同现模型的词同现网络构建方法，并对英汉词同现网络的特征进行了分析。近年随着神经网络技术的飞速发展，词向量表示性能得到了显著提升，方便了词相似度的计算，为构建基于相似性词同现网络奠定了理论基础。

为了研究相似性词同现网络技术及揭示藏语词同现网络的小世界效应和无标度特性，我们研究了藏文词同现网络构建方法，提出了一种基于相似度的藏文词同现网络构建方法，该方法以词为网络节点, 以相似的词间连边构造词同现网络。在大、中、小三类文档上建立了词同现网络，并分析了它们的统计特征，实验数据表明建立的藏文词同现网络都具有小世界效应和无标度特征。今后在该研究成果的基础上进一步研究藏文构件、字、词的超复杂网络构建技术及统计特征分析。

致谢

本项工作得到了国家自然科学基金资助项目（61866032,61966031), 青海省科技厅资助项目(2019-SF-129), “长江学者和创新团队发展计划”创新团队资助项目（IRT1068），青海省重点实验室项目（2013-Z-Y17、2014-Z-Y32、2015-Z-Y03），藏文信息处理与机器翻译重点实验室（2013-Y-17）资助。

参考文献

Steels L. 2000. *Language as a complex adaptive system.* Proceedings of Lecture Notes in Computer Science. Berlin: Springer-Verlag.

孙文俊，杜娟. 2010. 基于词同现网络与支持向量机的论文甄别. 现代情报, 2010(07):89-94.

才智杰, 孙茂松, 才让卓玛. 2018. 藏文字同现网络的小世界效应和无标度特性. 中文信息报,32(10):45-52.

才智杰,才让卓玛,孙茂松. 2020. 一种多基元联合训练的藏文词向量表示方法. 中文信息报,2020,34(5):44-49.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. *The Small World of Human Language.* Proceedings Biological Sciences, 268(1482):2261-2265.

Barabasi A L. 2002. *The New Science of Neworks.* Massachusetts,Persus Publishing.

梁伟，史玉明. 2012. 不同时期汉语散文的字同现网络之研究. 中国科学: 信息科学,42(7):831–842.

林枫，刘云，江钟立. 2012. 汉字网络的历时性模式探析. 复杂系统与复杂性科学,9(3):50-61.

刘知远，孙茂松. 2007. 汉语词同现网络的小世界效应和无标度特性. 中文信息学报,21(6):52-58.

Wei and Liang and YuMing. 2012. *Study on co-occurrence character networks from Chinese essays in different periods.* Science China Information Sciences,55(11):2417-2427.

Liang W and Wang Y and Shi Y. 2015. *Co-occurrence network analysis of Chinese and English poems.* Physic A: Statistical Mechanics and its Applications, 420:315-323.

Liang W and Shi Y and Tse C K. . *Comparison of co-occurrence networks of the Chinese and English languages.* Physic A: Statal Mechanics and its Applications, 388(23):4901-4909.

耿志杰, 王文鼐. 2012. 关键词同现网络结构研究. 情报杂志, 29(2):14-16.

余传明，周丹. 情感词汇共现网络的复杂网络特性分析. . 情报学报,29(5):906-914.

Liu R and Zhao H. 2014. *2011 International Conference on Management and Service Science-Word Co-Occurrence Network Analysis of Scientific Data Using NWB Tool.* IEEE 2011 International Conference on Management and Service Science (MASS 2011).

He B and Xu D. 2016. *An exploration on the word co-occurrence network of Chinese popular song titles..* International Conference on Natural Computation & Fuzzy Systems & Knowledge Discovery.

李亚星，王兆凯，冯旭鹏. 2016. 基于实时词共现网络的微博话题发现. 计算机应用,309(05):130-134.

Tsuya A and Sugawara Y and Tanaka A. 2014. *Do Cancer Patients Tweet? Examining the Twitter Use of Cancer Patients in Japan.* Journal of Medical Internet Research,16(5):e137.

刘知远，郑亚斌，孙茂松. 2008. 汉语依存句法网络的复杂网络性质. 复杂系统与复杂性科学,5(2):37-45.

Cancho R F I and Sole R V. 2001. *The Small World of Human Language.* Proceedings of the Royal Society of London Series B-Biological Sciences,268(1482):2261-2265.

Cancho R F I and Sole R V and Kohler R. 2004. *Patterns in Syntactic Dependency Networks.* Phys Rev E, 69(5):1915.

Motter A E and de Moura A P S and Lai Y C. 2002. *Topology of the Conceptual Network of Language.* Phys Rev E, 65(6):102.

才智杰，才让卓玛. 2011. 藏文自动分词系统的设计. 计算机工程与科学,33(5):151-154.

才智杰, 孙茂松, 才让卓玛. 2019. 藏文词向量相似度和相关性评测集构建. 中文信息学报,33(7):81-87,100.