文章编号:1001 - 9081(2010)07 - 1938 - 03

文本翻译索引的互文度量方法

姜 欣[1,2],姜 怡[1,2],方 森[3]
(1. 大连理工大学 电子信息与电气工程学部,辽宁 大连 116024; 2. 大连理工大学 外国语学院,辽宁 大连 116024;
3. 东北大学(秦皇岛分校) 电子信息系,河北 秦皇岛 066004)
(jiangxin1978@126.com)

摘 要:运用算法可更加科学地量化出翔实的显性互文线索,这对于追溯文本间的关联,理解和翻译文本都有着重要意义。以茶典籍文本为例,使用并比较了 4 种互文度量方法,即戴斯系数、匹配系数、全置信度和余弦,并给出用于文本辅助翻译的索引方法。文本互文度与互文度矩阵揭示了文本间的影响与关联。实验结果与性能分析表明余弦度量结果最好,基于互文性的文本翻译索引可为更加精确地理解和翻译相关文本提供有价值的参考。

关键词:度量方法; 互文性; 翻译索引; 茶典籍
中图分类号: TP305 **文献标志码**:A

Intertextuality measurement method of text translation index

JIANG Xin[1,2], JIANG Yi[1,2], FANG Miao[3]
(1. *Faulty of Electronic and Information Engineering, Dalian University of Technology, Dalian Liaoning* 116024, *China*;
2. *Foreign Languages School, Dalian University of Technology, Dalian Liaoning* 116024, *China*;
3. *Department of Electronic Information, Northeastern University at Qinhuangdao, Qinhuangdao Hebei* 066004, *China*)

Abstract: Computer algorithms can help produce scientific quantitative data to provide more reliable and precise manifest intertextual clues, which plays a significant role in clarifying the multilayer relationship among relevant texts, facilitating understanding and translating process. Using tea classics as case texts, this paper presented some algorithms to measure the intertextuality, namely, Dice coefficient, matching coefficient, full confidence and cosine, providing an index approach for text-oriented translation. The experimental results show that the cosine measure produces the best results, which offers valuable help to the accuracy and consistency in both source text comprehension and target version translation.

Key words: inmeasurement method; tertextuality; translation index; tea classics

0 引言

茶典籍文本呈现出超强互文性,分别体现在茶典籍文本的资源、论题、语篇、句法和词汇等不同层面,表现为显性的相似度和隐性的内涵共性。然而与此同时,它们也给文本的理解和翻译带来了诸多问题。梳理茶典籍文本间的互文关系是译者作为读者的首要步骤,是传神达意翻译的基础。英国著名的篇章语言学家哈蒂姆和梅森认为一系列的互文指涉必须被相互贯穿,并从后来所见文本中的符号指向前在的符号直至激起整个知识体系,从而找出其线索[1]。通过互文性分析,译者可以辨认出文本是如何与其他相关文本互相依靠,然后才能考虑如何传译原作中细微的互文标记,如何让读者真正领略到原文的风貌与文化内涵。

计算机作为人的中枢神经的延伸,作为信息存储和处理的新途径,亦可用于互文性分析。本项研究尝试根据几种相似度计算方法,如余弦系数(Cosine)、戴斯系数(Dice)、全置信度(Conf)和匹配系数(Match)来度量文本的互文性。相似度计算是知识表示以及信息检索的重要内容,意指对两个对象之间的相似程度的计算统计。相似度计算被广泛应用到文本分类、聚类、信息检索等多个领域,但是迄今尚未用于翻译中的互文指涉关系研究。鉴于文本不同层次的相似度是判断文间文内显性互文的重要依据之一,本文将从中国茶著的两部代表作,唐朝陆羽的《茶经》和清朝陆廷灿的《续茶经》入手,比较和运用这四种算法,以期更加全面、准确地追溯两个文本间多维互文指涉,为茶典籍的理解和翻译提供辅助的互文计算的优先列表。通过在茶经和续茶经的文本上的计算分析比较这四种度量的性能,并通过优先列表的评价表明互文计算在一定程度上为茶典籍译文本建构提供更具科学性的量化参考,为中华茶典籍的对外传播做出一种新的尝试。

1 文本互文性

1.1 互文性定义

互文性意指一个文本都是对其他文本的吸收和转换。研究表明任何文本的写作与阅读都有赖于此文本和彼文本的关系,文本的意义产生于和其他文本相互作用的过程中。“一个作者在写作自己的语篇时,会通过对另一(些)语篇的重复、模拟、借用、暗仿等,有意识地让其他语篇向本语篇产生扩散性的影响”[2]。

文本间的互文关系表现为语句之间的错综复杂的指涉,可以是字面上的,也可以是深层次的。因此英国语言学家诺曼费尔克劳在 Discourse and Social Change 中将互文分为显性互文和成构互文两个层次。显性互文是指一个语篇中标明的与其他语篇的互文关系。“在显性互文性中,其他语篇明显存在于所分析的语篇中,它们被语篇的表层特征,如引号,明

收稿日期:2009 - 12 - 14;**修回日期**:2010 - 02 - 09。 **基金项目**:国家自然科学基金资助项目(60673039);辽宁省教育厅 2009 年度高等学校科研项目计划;2008 年大连理工大学人文社科基金资助项目。

作者简介:姜欣(1959 -),女,江苏南京人,教授,主要研究方向:典籍英译; 姜怡(1959 -),女,江苏南京人,教授,主要研究方向:典籍英译; 方森(19 -),河南洛阳人,讲师,主要研究方向:机器翻译、语义信息检索。

确标示或暗示”[3]。我国学者罗选民也指出,显性互文指语篇的表层特征,手法有引用、模仿、糅杂、戏拟等;成构互文指过去和现在的体裁、规范、类型甚至主题在阅读的文本中发生的相互指涉关系[4]。

显性的互文表现为字面上的相似或相同,如术语、典故以及专有名词等的引用或仿拟;而成构或隐性的互文则处于更深的层次上,表现为无清楚标记的关联,只有通过语境、文化背景等才能得到正确的理解。因此,显性的互文性可以通过对文本的比较和一定的度量进行计算,其互文性度量亦可为隐性的互文理解提供一定的线索。

1.2　互文性度量方法

互文性度量通过一定的统计度量方法对两个相关语句进行度量。本文选取如下几种向量相似性度量[5]作为文本互文性的度量,度量定义如表 1 所示。

表 1　互文性度量的数学模型

互文性度量	数学模型	互文性度量	数学模型
戴斯系数	$\frac{2\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid}{\mid \boldsymbol{X}\mid+\mid \boldsymbol{Y}\mid}$	全置信度	$\frac{\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid}{\mathrm{Max}(\mid \boldsymbol{X}\mid,\mid \boldsymbol{Y}\mid)}$
匹配系数	$\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid$	余弦	$\frac{\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid}{\sqrt{\mid \boldsymbol{X}\mid * \mid \boldsymbol{Y}\mid}}$

把句子看做一个由字或词组成的向量,那么向量之间的统计值即是句子之间的互文性的度量。向量用 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 表示,$\mid \boldsymbol{X}\mid$,$\mid \boldsymbol{Y}\mid$ 分别是向量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的长度,即向量所含有的元素数目,$\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid$ 表示 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 公共维度,本文采用求解两个句子的最长公共子序列计算。最长公共子序列计算如下:

假设语句 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_m\}$ 和语句 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_n\}$ 的互文矩阵为:

$$\boldsymbol{D}=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

其中:

$$d_{ij}=\begin{cases}1, & x\text{ 的第 }i\text{ 元素与 }y\text{ 的第 }j\text{ 元素相同}\\ 0, & \text{其他}\end{cases}$$

根据这两个语句的互文矩阵 $d_{i,j}$,本文使用动态规划算法计算它们之间的最大匹配子序列。最大匹配子序列的计算是求两个语句的词语之间的最佳匹配,即在矩阵中寻找尽可能多的没有共同行和列的元素。最大匹配子序列的计算采用动态规划[6]算法实现,如下式所示:

$$\begin{cases} Intertextuality(\boldsymbol{D})=M_{m,n} \\ M_{i,j}=\max\{\max\limits_{i\leqslant k\leqslant j}\{d_{ik}+M_{i-1,k-1}\},\max\limits_{j\leqslant k\leqslant i}\{d_{kj}+M_{k-1,j-1}\}\}, \\ \qquad i\geqslant 2,j\geqslant 2 \\ M_{i,j}=\max\limits_{1\leqslant k\leqslant i,1\leqslant l\leqslant j}\{d_{kl}\}, \\ \qquad i=1\text{ 或 }j=1 \end{cases}$$

其中:$Intertextuality(\boldsymbol{D})$ 为相似度函数,$M_{m,n}$ 为动态规划目标函数。

例如“郭璞注:可以为羹饮,早采为茶,晚采为茗,一名舛。”⇔“今呼早取为茶,晚取为茗,或一曰荈,蜀人名之苦荼。”可以看作向量 $\boldsymbol{X}$ =(郭,璞,注,可,以,为,羹,饮,早,采,为,茶,晚,采,为,茗,一,名,舛)和向量 $\boldsymbol{Y}$ =(今,呼,早,取,为,茶,晚,取,为,茗,或,一,曰,荈,蜀,人,名,之,苦,荼)。公式$\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid$ 就是 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 共有子序列的元素数目,$\boldsymbol{X}\cap\boldsymbol{Y}$ =(为,晚,为,茗,一,名)。$\mid \boldsymbol{X}\cap\boldsymbol{Y}\mid=5$。而$\mid \boldsymbol{X}\mid$,$\mid \boldsymbol{Y}\mid$ 分别是向量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的元素数目,$\mid \boldsymbol{X}\mid=19$,$\mid \boldsymbol{Y}\mid=20$。

2　文本翻译索引

文本自动翻译是一件有意义但是非常困难的事情,中国古籍的翻译更是如此,然而我们可以利用计算机进行辅助翻译。文本的翻译索引是根据一定的度量标准自动选择出可供参考的原文语句。文本的互文性正是揭示文本间的影响与关联,可以为翻译的精确性和连贯性提供有益参考[7],不同文本间具有相互参照和指引作用,能构成互文,例如《茶经》和《续茶经》之间。同一文本内也可能包含了很多具有模因渊源的、可以相互参考的语句[8]。本文通过互文性的度量来寻找每一句的构成互文性的相关语句。

2.1　文本互文度

两个不同的文本可以计算互文度,假设原文(续茶经)有 n 句,历史文本(茶经)有 m 句。二者之间的互文度可以形成一个互文度矩阵 $\boldsymbol{A}_{(n\times m)}$:

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

其中每一个元素代表原文的一句与历史文本的一句之间的互文度,如 a_{ij} 代表原文第 i 句和历史文本的第 j 句之间的互文度。从互文度矩阵可以看出每一行代表原文的一句及其所有历史文本的索引,因此对每一行的元素按照互文度的高低进行排队,输出满足一定阈值的队列,即为原文语句的翻译索引。

2.2　矩阵的存储

从互文度矩阵可以看出,其中有很多元素的数值都是 0,因此,这两个矩阵都是稀疏矩阵。本文中采用数组和链表的形式来存储,如图 1 所示。

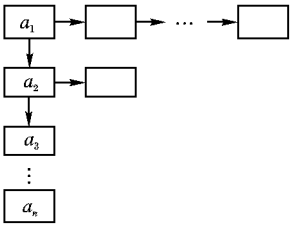

图 1　互文度矩阵的存储结构

图中的每一个节点代表一个语句,a_i 表示目标语句,即当前翻译的语句,右面的节点代表可参考的具有互文性的索引,或者是互文度不为 0 的参考节点。数值为 0 的节点被忽略。

3　实验与性能分析

3.1　实验结果

实验材料选择《茶经》和《续茶经》文本,分别计算互文度。本文实验的目的是检验统计度量和验证在统计度量下的翻译索引的性能。实验中选取其中满足一定阈值的部分结果进行比较,以检验统计度量和计算互文度。实验分成两个部分。

实验 1　文本间互文性度量检验。计算《茶经》与《续茶经》文本间的互文性,然后分别抽取句对子集进行比较,并对结果和计算准确率(accuracy)进行检查校对,准确率计算如下:

$Accuracy$ = 选取的结果中真正互文的数目 ×100%/选取

的结果中的互文数目

实验1结果如表2。

表2 各种度量模型的准确率

度量模型	句对数	准确率/%
戴斯系数	264	73.11
全置信度	256	71.09
匹配系数	206	61.65
余弦	269	75.84

实验2 翻译索引的有效性实验。由于互文性计算的一个直接目的就是帮助读者或者译者对古文进行理解和翻译，因此，互文性度量是否合适和有效还要针对每一个句子进行统计和分析。本文定义了另一个度量标准——Nbest准确率(Nbest_Accuracy)，即对一个句子寻找与之有互文关系的句子，并按照统计度量进行排序，Nbest就是取队列中的前N个。由于全文空间巨大，本文随机选取80个句子进行计算，并对准确率计算统计结果。

Nbest_Accuracy = Nbest中真正互文的数目×100%/集合中的句对数目

实验2的结果如图2所示，因为Nbest准确率在N取到10以后时变化不大，因此图中只给出列表中前10个句子的Nbest准确率的统计结果。

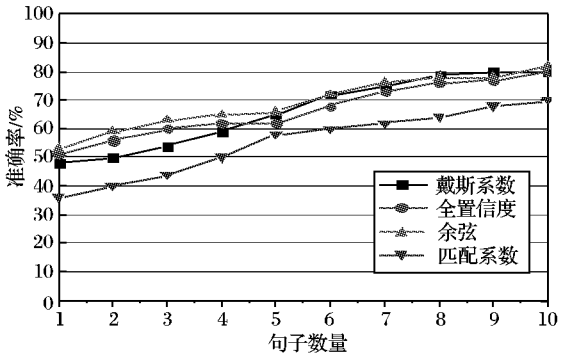

图2 翻译索引的有效性实验结果

3.2 性能分析

通过实验1可以看出，在这几种度量方法中，余弦度量的准确率较高，其次是戴斯系数和全置信度，最后是匹配系数。由此可以看出，余弦度量是较好的互文计算度量。

通过实验2可以看出，互文计算的结果的最好的序列Nbest，当N = 1时，余弦最好，大约在56%，匹配系数最差，大约在36%，戴斯系数和全置信度都在50%左右；而随着N的扩大，它们的准确率都在提高，当N = 10时，大约提高到约80%。而所有的句子中有10个句子没有与之构成互文的句子。由此可以看出，基于互文性的翻译索引可以为原文的理解与翻译提供有益的参考。

4 结语

本文利用戴斯系数、匹配系数、全置信度和余弦进行互文度量，并给出用于文本辅助翻译的索引方法。实验结果表明余弦度量结果较好，基于互文性的文本翻译索引为文本的理解和翻译提供有价值的帮助。通过在《茶经》和《续茶经》的文本上的计算，本文分析比较了这四种度量的性能，并通过优先列表的评价表明了互文计算在一定程度上为茶典籍揭示了文本间的影响与关联。

参考文献：

[1] HATIM B, MASON I. Discourse and the translator[M]. 上海：上海外语教育出版社，2001.

[2] 郑庆君. 互文型手机短信及其语篇特征探析[J]. 语言教学与研究，2007(5)：82－89.

[3] FAIRCLOUGH N. Discourse and social change[M]. Cambridge: Polity Press, 1992.

[4] 罗选民. 互文性与翻译[D]. 北京：清华大学，2006.

[5] MANNING C D, SCHUTZE H. Foundation of statistical natural language processing[EB/OL]. [2009－08－10]. http://www.sigmod.org/publications/sigmod-record/0209/b2.weikum.pdf.

[6] KRUSKAL J B. An overview of sequence comparison[C]// Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. MA: Addison-Wesley, 1983: 1－44.

[7] 姜欣，杨德宏.《续茶经》翻译中的互文关照[J]. 辽宁师范大学学报，2009，32(3)：92－94.

[8] 姜怡，姜欣. 中华茶典籍互文特质及其模因传播[J]. 农业考古，2009(6)：52－57.

(上接第1937页)

不敏感，实验表明基于全局K-means的谱聚类比基于K-means的谱聚类能得到好的聚类结果。

参考文献：

[1] HAN J W, KAMBER M. Data mining concept and techniques[M]. 范明，孟小峰，译. 北京：机械工业出版社，2001.

[2] HAMAD D, BIELA P. Introduction to spectral clustering[C]// Proceedings of 3rd International Conference on Information and Communication Technologies: From Theory to Applications. New York: IEEE, 2008: 1－6.

[3] MANOR L Z, PERONA P. Self-tuning spectral clustering [EB/OL]. [2009－09－10]. http://www.vision.caltech.edu/lihi/Publications/SelfTuningClustering.pdf.

[4] XIANG TAO, GONG SHAOGANG. Spectral clustering with eigenvector selection[J]. Pattern Recognition, 2008, 41(3): 1012－1029.

[5] 王玲，薄列峰，焦李成. 密度敏感的谱聚类[J]. 电子学报，2007，35(8)：1577－1581.

[6] 王玲，薄列峰，焦李成. 密度敏感的半监督谱聚类[J]. 软件学报，2007，18(10)：2412－2422.

[7] EKIN A, PANKANTI S, HAMPAPUR A. Initialization-independent spectral clustering with applications to automatic video analysis [C]// Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing. New York: IEEE, 2004: 641－644.

[8] 朱强生，何华灿，周延泉. 谱聚类算法对输入数据顺序的敏感性[J]. 计算机应用研究，2007，24(4)：62－64.

[9] LIKAS A, VLASSIS N, JVERBEEK J. The global K-means clustering algorithm[J]. Pattern Recognition, 2003, 36(2): 451－461.

[10] NG A, JORDAN M, WEISS Y. On spectral clustering: Analysis and an algorithm[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 897－856.

[11] WANG WEINA, ZHANG YUNJIE, LI YI, *et al*. The global fuzzy c-means clustering algorithm[C]// Proceedings of the 6th World Congress on Intelligent Control and Automation, New York: IEEE. 2006.

[12] 汪中，刘贵全，陈恩红. 基于模糊K-harmonic means的谱聚类算法[J]. 智能系统学报，2009，4(2)：95－99.