面向知识挖掘的平行句法语料库构建研究

数字人文视角下的史部典籍信息组织

王东波◎著

非外借

南京大学出版社

面向知识挖掘的平行句法语料库构建研究

数字人文视角下的史部典籍信息组织

王东波◎著

本书出版获得以下研究项目的资助：
国家自然科学基金青年项目（71303120）：基于CSSCI的句法级汉英平行语料库及知识挖掘研究
国家自然科学基金面上项目（71673143）：基于典籍引得的句法级汉英平行语料库构建及人文计算研究

南京大学出版社

图书在版编目(CIP)数据

面向知识挖掘的平行句法语料库构建研究 / 王东波著. —南京：南京大学出版社，2019.12
ISBN 978-7-305-18811-4

Ⅰ. ①面… Ⅱ. ①王 Ⅲ. ①汉语—句法—研究 ②英语—句法—研究 Ⅳ. ①H146.3 ②H314.3

中国版本图书馆 CIP 数据核字(2017)第 132302 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
出 版 人 金鑫荣

书　　名 面向知识挖掘的平行句法语料库构建研究
著　　者 王东波
责任编辑 陈亚明　　编辑热线 025-83592401

照　　排 南京理工大学资产经营有限公司
印　　刷 江苏凤凰数码印务有限公司
开　　本 787×960 1/16 印张 25.75 字数 350 千
版　　次 2019 年 12 月第 1 版 2019 年 12 月第 1 次印刷
ISBN 978-7-305-18811-4
定　　价 80.00 元

网　　址：http://www.njupco.com
官方微博：http://weibo.com/njupco
官方微信号：njupress
销售咨询热线：025-83594756

FOREWORD 前言

唐太宗李世民说过:“以史为镜,可以知兴替。”史书中不仅蕴含了五彩斑斓的百科知识,而且可以根据以往有效地预测未来。中国有着悠久的历史,而历代的史学典籍则是这悠久历史的最好记录者。随着信息技术的迅猛发展,如何把数字人文的理念、方法和理论结合人工智能下的自然语言处理的技术、方法和模型有效地融入历史典籍的知识挖掘中,从而实现快速而精准地从典籍中有效挖掘相应的历史知识成为当下一个研究的热点和趋势。

在上述这一大的背景下,我们基于自然语言处理中的句子对齐、自动分词、词性标注、实体识别和短语结构识别等技术,面向《左传》《战国策》《史记》《汉书》《后汉书》《三国志》的典籍文本和对应白话文本,构建了古白平行语料库,并结合传统机器学习和深度学习的系列模型,从数字人文的视角完成了对史部典籍的知识组织。在对史部典籍知识组织的基础上,对典籍中所蕴含的类别和涵盖的相应问题进行了知识的挖掘和探究。本书的具体研究内容如下:

通过对比的方法,借助 CiteSpace 的共被引聚类功能进一步探究国际数字人文研究中计算机和人文学科的具体研究内容。确定了国际数字人文研究中计算机学科的 6 个类群和 2 个研究主题,人

文学科中的5个类群和2个研究主题。

结合古文和白话文的特征，本书基于支持向量机模型(SVM)和LSTM-CRF模型，将句子对齐视为双语候选句对中“对齐句对”与“非对齐句对”分类问题以及标注问题，结合实验文本特点提取较为有效的特征，基于“整体分类”与“序列标注”两种理念实现句子对齐并围绕特征选取与组合展开讨论。

结合自动分词的相应知识，本书基于Bi-LSTM-CRF模型和Bi-LSTM模型，构建了古文和白话文自动分词模型。在构建的古汉语自动分词模型中，调和平均值为95.87%的Bi-LSTM-CRF训练模型可作为最佳模型；在构建的白话文自动分词模型中，调和平均值为92.86%的Bi-LSTM的训练模型可作为白话文自动分词的最佳模型。

针对古文和白话文词性标注的任务，本书基于Bi-LSTM-CRF模型，构建了古文和白话文词性自动标注模型，对古白平行语料进行了词性标注。在模型的构建过程中，通过对测试语料的十折交叉验证，证明了基于Bi-LSTM-CRF模型构建的古文和白话文词性自动标注模型相较于基于Bi-LSTM模型构建的古文和白话文词性自动标注模型效果更优。

对于古文实体抽取这一研究，本书选择条件随机场模型、Bi-LSTM模型、Bi-LSTM-CRF模型这三种模型，基于《史记》预先处理过的语料训练出人名实体标注的模型，然后利用训练出的模型，对没有标注的部分史书，包括《三国志》《汉书》《后汉书》《左传》《战国策》，完成人名实体的标注。

基于已有的短语结构知识和特征，本书通过CRF、Bi-LSTM以及Bi-LSTM-CRF三种模型，以清华汉语树库和《史记》作为原始语

料，以介宾短语结构为切入点，构建了古白介宾结构自动识别模型，并完成了对《三国志》《汉书》《后汉书》《左传》《战国策》等古白典籍中所包含的介宾短语结构的识别。

在上述所构建的古白平行句子对基础上，结合所形成的句法层面的传统机器学习和深度学习模型，针对典籍中所蕴含的类别知识和相应问题信息，本书对典籍问题的自动分类和典籍知识问答这两类研究进行了探究，并搭建了典籍问题自动分类模型和基于典籍知识图谱的自动问答平台。

无论是数据的标注，还是模型的构建，抑或是实验结果的描述，从工作量上看，本研究均需要投入大量的时间和人力。在本书的完成过程中，有如下人员参与了数据标注、模型构建和实验结果描述的工作：高正、李志豪、刘忠诚、丁可、王宗昊、高瑞卿、唐梦嘉、刘肖、王璐璐、闫文浩、宋旭雯、薛嘉楠、张琪、梁继文、叶文豪、胡昊天。对于上述参与完成本书的各位人员，再次表示深深的感谢，没有你们的参与，就不会有这本书的出版。由于本书是利用数字人文的理念和方法进行的一次尝试性探究，在历史典籍数据的标注、模型的构建和结果的分析与描述上定会存在不当之处，敬请数字人文、历史学、计算机科学、语言学和文献学的研究者、爱好者和对本研究感兴趣的读者包涵、谅解、指正和批评，本书的作者定会逐一修改，并万分感谢。

王东波

2019 年于南京

CONTENTS 目录

第 1 章　引言 …… 1

1.1　本研究的意义 …… 2

1.2　本研究的现状及发展动态分析 …… 2

1.3　研究方法 …… 10

1.4　本书的特色与创新之处 …… 10

1.5　小结 …… 11

第 2 章　定量对比视角下的国内外数字人文研究进展 …… 12

2.1　国内国际数字人文研究学科参与度与机构合作模式差异 …… 14

2.2　国内国际数字人文研究差异内容分析 …… 20

2.3　小结 …… 30

第 3 章　基于史部典籍的古白句子对齐及分析研究 …… 32

3.1　相应研究梳理 …… 32

3.2　模型介绍及对齐模型构建 …… 39

3.3　最优模型的应用 …… 59

3.4　小结 …… 62

第 4 章　词汇级的古白平行语料库构建及分析研究 …… 64
4.1　相应研究梳理 …… 65
4.2　自动分词研究概述 …… 71
4.3　史部典籍古白自动分词模型构建 …… 79
4.4　史书类古白词汇分布分析 …… 94
4.5　小结 …… 116

第 5 章　词性级的古白平行语料库构建及分析研究 …… 118
5.1　相关研究梳理 …… 119
5.2　词性标注模型构建 …… 127
5.3　所有史书类古白语言词性分布分析 …… 143
5.4　小结 …… 150

第 6 章　实体级的古白平行语料库构建及分析研究 …… 152
6.1　命名实体识别研究现状 …… 155
6.2　实证研究 …… 168
6.3　所有史书类古白语言姓名分布分析 …… 187
6.4　小结 …… 195

第 7 章　句法级的古白平行语料库构建及分析研究 …… 197
7.1　相关研究概述 …… 198
7.2　介宾结构自动识别研究 …… 207
7.3　三种模型的训练过程及结果分析 …… 218
7.4　古白史书的介宾结构识别及分析 …… 226
7.5　小结 …… 235

第 8 章 基于语料库和模型的典籍类别知识挖掘研究 …… 237
8.1 引言 …… 237
8.2 相关研究状况 …… 238
8.3 模型介绍及对齐模型构建 …… 242
8.4 基于典籍问句语料库的特征选择及分类实验 …… 246
8.5 小结 …… 258

第 9 章 基于语料库和模型的典籍语义知识挖掘研究 …… 259
9.1 功能设计 …… 260
9.2 实现算法 …… 263
9.3 典籍问答系统功能展示 …… 267
9.4 小结 …… 269

附件一 史部典籍文白对齐句子 …… 270
附件二 史部典籍分词语料 …… 292
附件三 史部典籍词性标注语料 …… 316
附件四 史部典籍人名标注语料 …… 339
附件五 史部典籍浅层句法标注语料 …… 362

参考文献 …… 386

第1章

引　言

中国历史典籍浩如烟海，在历史发展的长河中赋予了中华民族特有的个性和民族身份感。对历史典籍中的优秀文化进行深度挖掘，把这些优秀的历史文化介绍给世界，是促进世界对中国的了解和接受，实现中外文化交流，达到世界文化融合的重要途径之一。习近平总书记在“全国宣传思想工作会议”上强调，“讲清楚中华优秀传统文化是中华民族的突出优势，是我们最深厚的文化软实力。”在数字人文日渐兴起的大背景下，研究如何从典籍文本中挖掘出新的知识、统计分析典籍中古白词汇的差异、计算出古白文体风格，对提升中国文化的软实力也具有重要的意义。由汉语典籍原文文本及其平行对应的现代汉语译语文本构成的典籍平行语料库不仅为古白典籍的检索及呈现提供了最基本的素材，而且也是相关数字人文得以进行的基础。但目前古白平行的典籍平行语料库一方面数据量比较小，一般均是由某一部典籍构成，缺乏系统性、历时性；另一方面古白典籍平行语料库的平行单位和标注层级相对单一和浅显，一般到段对齐和词性标注，并且对基于古白典籍平行语料库的利用缺乏数字人文的探究。在上述背景下，本研究基于典籍的古文和现代汉语数据，通过句子对齐算法、分词、词性、实体和短语等句法级的相应技术、方法和知识，构建句子层级的典籍古白平行语料库，并在构建的语料库上进行知识挖掘的探究。

1.1 本研究的意义

本研究的意义主要体现在典籍句法级古白平行语料库构建方法、典籍句法级古白平行语料库、基于平行语料库的数字人文、中华文化走出去及话语体系的构建等方面。

第一，典籍句法级古白平行语料库构建过程中涉及各种技术手段和模型，本研究在技术手段方面的改进、机器学习模型特征选取上的创新对语言学、文献学和情报学的语料构建和非结构化知识的挖掘具有借鉴作用。

第二，在基于平行语料库的数字人文探究过程中，从古白跨语言、跨文化的角度计算了词汇、篇章和版本各个层面的差异，从古白双语计算的角度丰富数字人文的方式方法，促进数字人文理论的构建，有助于中国的数字人文尤其是针对古代文献的数字人文在国际上争取话语权。

第三，所构建的句法级古白平行语料库不仅可以为跨语言检索衍生出专业性更强的古文与白话文双语词典，而且在一定程度上有助于跨语言检索从句法级语言资源中获取相应的语义知识，从而为语义网的探究提供相应的知识。同时，本研究构建的句法级古白平行语料库在一定程度上也可以为古白语言信息处理任务中的领域本体构建提供相应的支持。

第四，在中华文化走出去及话语体系构建方面，典籍句法级古白平行语料库的构建及数字人文将增强中华传统文化思想精髓对外译介的传播有效性，并有利于学术话语体系构建中传统学术话语资源的补充。

1.2 本研究的现状及发展动态分析

围绕着典籍句法级古白平行语料库的构建及数字人文这一主题，国

内外的研究综述主要从平行语料库的构建、句法分析这两个大的方面展开：

(1) 平行语料库构建的研究

平行语料库是由两种或两种以上语言的语料构成的。目前的语料库构建中，单语语料库相对发展较快，在不同标注层级上，有代表性的单语语料库如下：国外有美国布朗大学的 Brown 语料库(100 万词的美国英语)、柯林斯-伯明翰大学的 COBUILD 国际语料库(2 亿词的英语)、美国宾州大学为句法分析而设计的树库(Penn Treebank)等；国内也有富士通研究开发中心和北京大学计算语言学研究所等单位研制的基于《人民日报》的“汉语词性标注语料库”①，国家语委和几家高校、科研机构在“863”著作支持下正在建设的 15 亿字超大规模平衡语料库，台湾“中研院”200 万词次带有词性标注的汉语平衡语料库以及在“973”著作支持下清华大学建设的汉语句法树库(以下简称 973 树库)②等代表性的汉语单语语料库。和单语语料库相比，双语平行语料库的建设起步较晚，数量也较少。最著名的平行语料库当属加拿大的议会会议录(Canada Hansards)，该会议录同时用英、法两种语言记录而成③，许多最初的平行语料库的研究都是在该语料库的基础上进行的。此外，20 世纪 90 年代建立的英语-挪威语双语语料库、英语-意大利语双语语料库，以及英国曼彻斯特大学科技学院翻译研究中心的翻译语料库(简称 TEC)等也都很著名④。而包含语言最多的平行语料库，是圣经语料库(The Bible Corpus)，它由马里兰大学的 Resnik 等人构建，包含了 9 种语言(英语、法语、丹麦语、芬兰

① 杨惠中等.语料库语言学导论[M].上海：上海外语教育出版社，2002.

② 吕雅娟.基于双语语料库对齐的翻译知识自动获取技术研究[D].哈尔滨工业大学，2003.

③ Gale W A, Church K W. Identifying Word Correspondences in Parallel Texts[J]. HLT, 1991.

④ Stawart D T C. Corpora in Translation Studies[M]. Dicho, 2012.

语、希腊语、瑞典语、拉丁语、西班牙语、越南语)[①]。

近几年来,平行语料库的研究价值越来越得到国内学者的关注。许多大学和研究机构开始进行汉外平行语料库的建设。比较著名的是香港科技大学的英语和广东话双语语料库(简称 HKUST),其主要内容是香港立法委员会的会议记录[②]。中国大陆已建立的平行语料库有:北京大学计算语言研究所、清华大学智能技术国家重点实验室和中国科学院计算技术研究所共同开发的"面向新闻领域的古白翻译系统",该语料库已收集到的古白对照语料有中文约 2 000 万字,英文约 1 000 万单词[③];北京外国语大学中国外语教育研究中心正在建设的 2 000 万字汉日平行语料库和 3 000 万字/词的古白平行语料库[④]。此外,哈尔滨工业大学[⑤]、东北大学[⑥]、台湾"中研院"、南京师范大学等也都在进行双语平行语料库的建设工作。但这些双语平行语料库的规模多数在 10 万句对以下,且加工深度仅限于句子对齐。

在以典籍为代表的古汉语语料库构建中,单语语料库相对较少,并且标注的层级也相对比较简单。郁默[⑦]对以《十三经》为主的先秦文献进行了分词探究。邱冰和皇甫娟[⑧]融合字的互信息(Mutual Information)和词

① Resnik P, Olsen M B, Diab M. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues"[J]. Computers & the Humanities, 1999, 33(1-2):129-153(25).

② Wu D. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria[J]. ACL-94, 1994:80-7.

③ 黄俊红,黄萍,范云.专门用途语类翻译平行语料库研究述评[J].重庆大学学报(社会科学版),2004,10(6):91-94.

④ 王克非.英汉/古白语句对应的语料库考察[J].外语教学与研究,2003,35(6):11-17+82.

⑤ Yang M. Y. A Research on Bilingual Dictionary Based Sentence Alignment for Chinese English Parallel Corpus[J]. High Technology Letters, 2002, 8(1):8-11.

⑥ 吕学强,李清隐,陈文亮,等.古白法律文献的子条级自动索引和对齐[J].中文信息学报,2002(04):52-59.

⑦ 郁默.台湾"中研院"汉籍全文资料库[J].中国典籍与文化,1998(03):110-115.

⑧ 邱冰,皇甫娟.基于中文信息处理的古代汉语分词研究[J].微计算机信息,2008(24):100-102.

汇的频率知识在先秦语料上进行了分词实验。徐润华等[①]以《左传》为例，提出了一种利用古籍注疏文献中的词汇语义知识，在文献和注疏自动对齐的基础上进行的古文献分词新方法。梁社会等[②]分别使用 CRF 模型和注疏文献方法进行《孟子》分词研究，发现两种方法在《孟子》中均取得了理想的分词效果。以现代汉语为基线训练语料，留金腾等[③]在人工校对的基础上，对《淮南子》完成了分词标注。化振红[④]在古汉语语料库构建若干问题的论述中，提出了古汉语分词规范的整体框架。王嘉灵[⑤]对《汉书》进行了详尽的分词研究，发现地名表、人名表加上注疏词表能达到最好的分词效果。同时她还使用 CRF 模型进行自动分词，添加了声、韵等语言特征，自动分词 F 值达到 94.4%。黄建年[⑥]探讨了计算机技术在农业古籍断句标点、分词标引中的应用，构建了农业古籍断句标点、分词标引的原型系统。石民等[⑦]实现了一种基于 CRF 模型的先秦汉语分词及词性标注一体化的系统。朱晓和金力[⑧]以《明史》为语料对象，验证了条件随机场在无边图模型、完全图模型以及嵌套图模型三种模型上的性能，并得出完全图模型和嵌套图模型较为突出的结论。于丽丽等[⑨]研

① 徐润华，陈小荷.一种利用注疏的《左传》分词新方法[J].中文信息学报，2012(02)：13－17，45.

② 梁社会，陈小荷.先秦文献《孟子》自动分词方法研究[J].南京师范大学文学院学报，2013(03)：175－182.

③ 留金腾，宋彦，夏飞.上古汉语分词及词性标注语料库的构建——以《淮南子》为范例[J].中文信息学报，2013(06)：6－15，81.

④ 化振红.深加工中古汉语语料库建设的若干问题[J].西南大学学报(社会科学版)，2014(03)：136－142，184.

⑤ 王嘉灵.以《汉书》为例的中古汉语自动分词[D].南京师范大学，2014.

⑥ 黄建年.农业古籍的计算机断句标点与分词标引研究[D].南京农业大学，2009.

⑦ 石民，李斌，陈小荷.基于 CRF 的先秦汉语分词标注一体化研究[J].中文信息学报，2010(02)：39－45.

⑧ 朱晓，金力.条件随机场图模型在《明史》词性标注研究中的应用效果探索[J].复旦学报(自然科学版)，2014(03)：297－304.

⑨ 于丽丽，丁德鑫，曲维光，等.基于条件随机场的古汉语词义消歧研究[J].微电子学与计算机，2009(10)：45－48.

究比较了 CRF 与最大熵和朴素贝叶斯统计模型，发现 CRF 进行古汉语词义消歧的实验效果最好。董志翘[①]在论述中古汉语研究型语料库构建的过程中，阐述了中古阶段词汇语义的关系体系。张颖杰等[②]提出了一种新颖的先秦汉语词义标注方法，利用 SVM 对《左传》进行了半指导的词性标注实验，平均正确率达到 67%。该研究为普遍缺乏训练语料的古籍文本词义标注提供了一个可行的研究思路。在对于古汉语词义本身的研究中，刘浏等[③]从词语的时代意义角度，提出了一种先秦词语时代特征的自动获取方法，并在此基础上提出了利用朴素贝叶斯模型进行文献时代自动判定的方法。梁社会等[④]从修辞学角度入手，以《论语》《孟子》为例，详细分析了先秦汉语排比句的特点，并设计了自动识别排比句的算法。许超等[⑤]利用 Pajek 软件，对《左传》中提取的人物和事件建立了社会网络，并以此为基础，对该时期的社会网络关系进行了定性定量的探索性研究。该研究为我们利用古文献探究古代社会关系提供了一个新的思路。

与典籍单语语料库相比，与典籍相关的平行语料加工和挖掘技术研究相对比较匮乏。马创新等[⑥]以《论语》为例，阐述了构建古籍与其注疏文献对齐语料库的必要性。马创新等[⑦]使用 XML 语言描述了《论语》及其注疏文献对齐语料库中的知识，提出了一种充分利用计算语言学研究

① 董志翘.为中古汉语研究夯实基础——"中古汉语研究型语料库"建设琐议[J].燕山大学学报(哲学社会科学版),2011(01):1-6.

② 张颖杰,李斌,陈家骏,等.基于词典信息的先秦汉语全文词义标注方法研究[J].中文信息学报,2012(03):65-71,103.

③ 刘浏,李斌,曲维光,等.先秦词汇的时代特征自动获取及文献时代的自动判定[J].中文信息学报,2013(05):107-113.

④ 梁社会,陈小荷,刘浏.先秦汉语排比句自动识别研究——以《孟子》《论语》中的排比句自动识别为例[J].计算机工程与应用,2013(19):222-226.

⑤ 许超,陈小荷.《左传》中的春秋社会网络分析[J].南京师范大学文学院学报,2014(1):179-184.

⑥ 马创新,陈小荷,曲维光,等.《论语》与其注疏文献对齐语料库的构建[J].现代教育技术,2012(7):109-113.

⑦ 马创新,陈小荷.基于 XML 的《论语》与其注疏文献对齐语料库的知识表示[J].图书情报知识,2013(1):107-113.

古籍文献的新思路。

已有的单语语料库构建及相应的技术为典籍古白双语语料库构建提供了相应的借鉴思路、方法和策略，从古白双语语料的标注深度上分析，随着知识挖掘技术的快速发展，对古白语料进行句法标注逐步成为一种趋势。

(2) 语法分析的相关研究

Bar-Hillel[①] 提出了范畴语法，任何词都可以根据它在句子中的功能归入一定的句法类型，范畴语法在描写英语方面获得了满意的效果。秦莉娟和周昌乐[②]采用基本词库加扩展生成的思想构建了面向范畴语法分析的汉语词库。周明等[③]采用依存文法，提出一种基于依存文法的融合语料库、规则方法和统计方法的汉语分析模型 CRSP。刘贵全等[④]提出采用最大熵模型实现中文依存语法的分析。

基于规则的句法分析方法主要有 Early 算法、LR 分析算法、Chart 算法和 GLR(富田胜)算法等，其中使用最广泛的是 Chart 算法和 GLR 算法。基于规则的方法在处理大规模真实文本时，会存在语法规则覆盖度有限、系统可迁移性差等缺陷。随着大规模标注树库的建立，基于统计学习模型的句法分析方法开始兴起，最典型的就是 PCFG(Probabilistic Context Free Grammar)，为了突破 PCFG 所做的独立性假设，出现了词汇化 PCFG 方法。与英文句法分析相比，中文句法分析的研究相对较晚，但在上述算法的基础上，中文句法分析的研究者们还是提出了很多改进与优化的算法。朱胜火等[⑤]提出并实现了一种有效的概率上下文无关文

① Bar-Hillel Y. A Quasi-Arithmetical Notation for Syntactic Description[J]. Language, 1953, 29(1):47-58.

② 秦莉娟，周昌乐.面向范畴语法分析的汉语词库的构造及实现[J].中文信息学报，2001，03.

③ 周明，黄昌宁，张敏，等.统计与规则并举的汉语句法分析模型[J].计算机研究与发展，1994，02

④ 刘贵全，曾宇斌.基于最大熵模型的汉语依存分析[J].计算机工程，2006，11.

⑤ 朱胜火，周明，刘昕，等.一种有效的概率上下文无关文法分析算法[J].软件学报，1998，08.

法 SCFG 的分析算法，对 GLR 分析表加以改造。周强[①]提出了一种分阶段构造汉语树库的设想。杨开城[②]提出了一种基于句法语义特征的汉语句法分析器。林颖等[③]提出了一个基于统计模型的自顶向下的汉语句法分析器。熊德意等[④]提出从树库中获取丰富的语言信息，以改善句法分析模型的性能。刘贵全等[⑤]采用最大熵模型实现了中文依存语法的分析。李幸等[⑥]提出了一种新的面向汉语长句的层次化句法分析方法。曹海龙等[⑦]把句法分析分解为分词及词性标注、短语识别两个部分。

为了解决句法匹配过程中的数据稀疏问题，已有研究融入了词汇搭配的相关知识。关于搭配的相关研究主要内容如下。20 世纪 70 年代，Jones 和 Sinclair[⑧] 发表了第一个基于语料库的词语搭配研究报告。Choueka 等[⑨]较早展开了搭配获取方面的研究，该文把搭配定义为重复出现的紧邻的词构成的序列。Church 和 Hanks[⑩] 把搭配定义为相互联系的词对，使用信息论中的互信息为指标来评价两个词的结合能力。

① 周强.汉语匹配算法的实现[C]//陈力为，袁琦编.语言工程.北京：清华大学出版社，1997.

② 杨开城.一种基于句法语义特征的汉语句法分析器[J].中文信息学报，2000，03.

③ 林颖，史晓东，郭锋.一种基于概率上下文无关文法的汉语句法分析[J].中文信息学报，2006，02.

④ 熊德意，刘群，林守勋.融合丰富的语言知识的汉语统计句法分析[J].中文信息学报，2005，03.

⑤ 刘贵全，曾宇斌.基于最大熵模型的汉语依存分析[J].计算机工程，2006(11)：216－218.

⑥ 李幸，宗成庆.引入标点处理的层次化汉语长句句法分析方法[J].中文信息学报，2006(04)：8－15.

⑦ 曹海龙，赵铁军，李生.基于词汇化模型的汉语句法分析[J].电子与信息学报，2007(09)：2082－2085.

⑧ Jones S，Sinclair J.English Lexical Collocations：A Study in Computational Linguistics[J].Cahiers-de-Lexicologie，1974，24(1)：15－61.

⑨ Choueka Y，Klein T and Neuwitz E. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus[J].Journal for Literary and Linguistic Computing，1983(4)：34－38.

⑩ Church，K W，Hanks P. Word Association Norms，Mutual Information and Lexicography[J].Computational Linguistics，1990，16(1)：22－29.