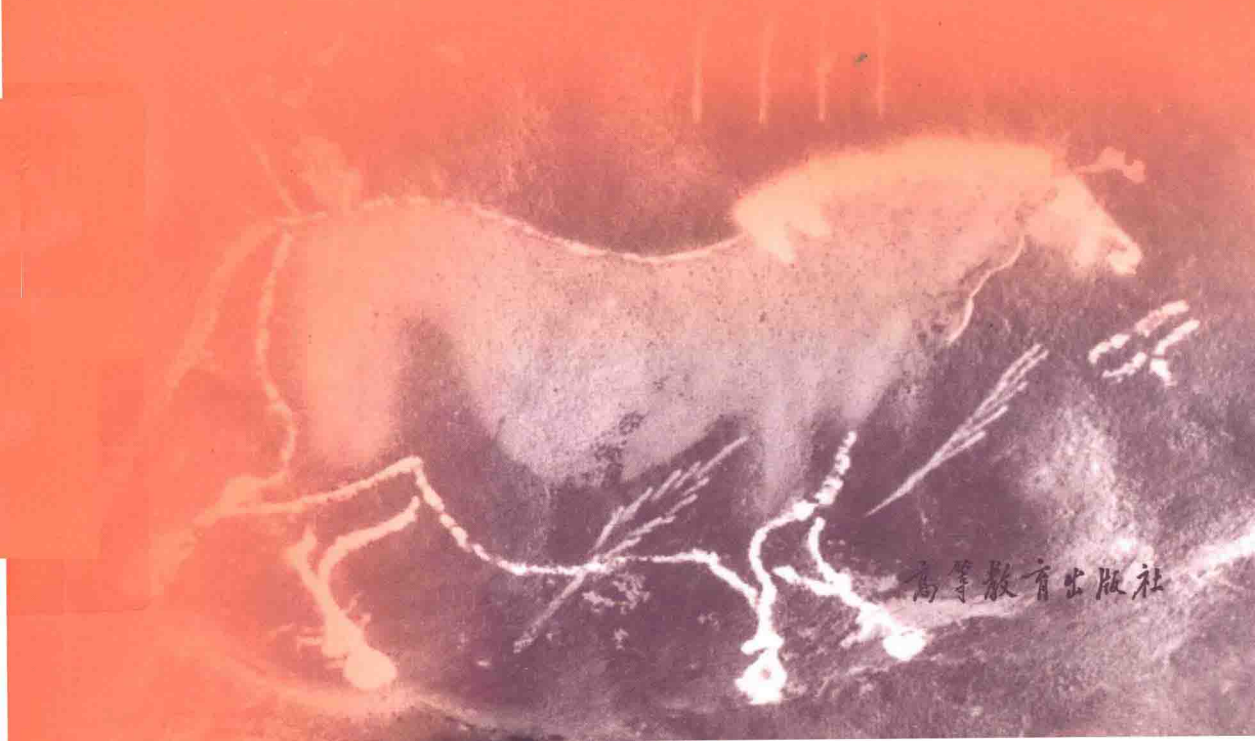
胡开宝 主编

Vol.1

Corpora & Intercultural Studies

语料库与跨文化研究

（第1辑）



高等教育出版社

● 胡开宝 主编

Vol.1

Corpora & Intercultural Studies

语料库与跨文化研究

（第1辑）

高等教育出版社·北京

图书在版编目（CIP）数据

语料库与跨文化研究．第1辑 / 胡开宝主编．-- 北京 ：高等教育出版社，2017.11
ISBN 978-7-04-048791-6

Ⅰ．①语… Ⅱ．①胡… Ⅲ．①语料库－研究
Ⅳ．①H0

中国版本图书馆CIP数据核字(2017)第273557号

策划编辑 常少华　　责任编辑 常少华　　封面设计 王 鹏
版式设计 王东岗　　责任校对 巩 婕　　责任印制 韩 刚

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京东君印刷有限公司
开　　本 787mm×1092mm 1/16
印　　张 10.75
字　　数 227千字
购书热线 010-58581118
咨询电话 400-810-0598

网　　址 http://www.hep.edu.cn
http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
http://www.hepmall.com
http://www.hepmall.cn

版　　次 2017年11月第1版
印　　次 2017年11月第1次印刷
定　　价 29.00元

物 料 号 48791-00

序

半个多世纪前，Brown语料库的创建者Nelson Francis第一次用“Corpus”一词，指代大量语言材料的集成。自此，语料库开始进入西方学者的视野，而现代语料库语言学的真正发轫却始于20世纪80年代，其学科基础源于John Sinclair教授提出的一系列重要语言学思想。在英国，实证主义传统为语料库语言学研究提供了肥沃的土壤，而日新月异的计算机信息技术有力推动了语料库语言学的迅猛发展。语料库语言学研究阵营日益壮大，一大批优秀的语料库语言学家登上语言学研究的历史舞台，提出了一系列具有影响力的学术思想，语料库语言学也成为一门重要的语言学科，受到广泛关注。

语料库语言学领域内一个重要的研究方向是基于语料库的应用研究。在英国的兰卡斯特大学，Tony McEnery教授创建了UCREL语料库语言学研究中心，该中心致力于将语料库应用到社会科学研究之中；在曼彻斯特大学，Mona Baker教授建设了世界上第一个翻译语料库，开创了基于语料库的翻译学研究范式。另外，通过检索最近5年发表在语料库语言学国际期刊上的文章后发现，大多数的文章是有关语料库的应用性研究，涉及的交叉学科领域之广、内容之丰富，不一而足。

20世纪80年代初，我国建成了第一个大型的电子语料库，即“交大科技英语语料库”（JDEST），它属于世界上第一批电子语料库。建设JDEST的目的就是要服务于当时国内的外语教学。经过三十多年的发展，国内业已建成了许多不同类型的语料库，语料库研究也呈现出一片欣欣向荣的景象。目前，语料库已逐渐应用到社会学、语言学、翻译、文学、教学、医学、人工智能乃至自然科学研究之中。

在这样的背景下，我们创办《语料库与跨文化研究》，旨在搭建学术平台，发表关于语料库应用的最新研究成果，以期推进语料库在人文社会科学和自然科学中的应用。《语料库与跨文化研究》由“中国语料库与跨文化研究论坛”主办，秉持兼容并蓄的理念，大力倡导语料库与其他各个学科的交叉性研究，常设“语料库语言学研究”“语料库翻译学研究”和“语料库与文化研究”三个主要栏目。是否能够实现创办初衷，有赖于学术界同仁的不吝匡正与精诚合作。

目　录

CONTENTS

语料库系统的评测

冯志伟 杭州师范大学

提　要：　本文从自然语言处理系统评测的角度，讨论语料库系统的评测问题，主要有：语料库的类型、语料库的元数据、语料库自动切词和自动标注的评测原则与正确率的计算公式。

关键词：　语料库；评测；元数据；自动切词；自动标注

评测是推动自然语言处理研究发展的一种重要手段（冯志伟 1997）。本文参照有关国际标准和国家标准，从标准化和规范化的角度，来讨论语料库的评测原则和方法。我们的这些意见仅供从事语料库研究的同行参考，只有推荐性，没有强制性（冯志伟 2010）。

语料库系统的评测属于自然语言处理系统评测（例如，语音合成和文语转换系统评测、语音识别系统评测、机器翻译系统评测、语料库系统评测）的一部分。自然语言处理系统的评测标准（assessment norms）是用于评测的标准体系，包括评测内容、评价指标、评测方法和文件格式等。在评测的时候，我们要依据一定的技术指标体系和有关规范，采用一定的方法和程序，对于自然语言信息处理系统及其组成要素的功能、特性和运行效果进行评价和检测。

自然语言处理系统的评测应遵守如下7个原则：

（1）公平公正的原则：评测应努力做到公平、公正。

（2）遵循标准的原则：评测应遵循国际标准、国家标准和相关语言文字规范。

（3）人机结合的原则：在当前条件下，基本上应以人工评测为主，辅之以机器自动评测。

（4）区别对待的原则：评测应针对不同语言信息处理系统和用户类型的特点，区别对待。

（5）灵活柔性的原则：语言文字具有一定的灵活性和柔性，并非处处都是界限分明、非此即彼的，在遵循标准的前提下，有时可以容许两种或多种可能的结果并存。

（6）可操作性的原则：评测应当是可以操作的，评测时，应当具体地说明评测的方法、步骤和评分方式。

在自然语言处理系统的评测中，有两种不同的评测方法，一种是黑箱评测（black box assessment），一种是白箱评测（glass box assessment）。

在进行黑箱评测时，不关心自然语言信息处理系统的内部机制和组成结构，主要根据系统的输入数据与输出结果进行判断。黑箱评测有助于了解自然语言信息处理系统外在的总体性能，又叫作“外在评测”（extrinsic assessment）。

在进行白箱评测时，需要对自然语言信息处理系统的内部机制分别进行分析，逐一评测系统的各个组成部分的性能。白箱评测可以针对信息处理系统的各个组成部分分别进行，对于不同的部分准备不同的测试数据，从而判断所出现的错误是在哪一个部分造成的，这样就可以为规则的调整和算法的改进提供可靠的数据。白箱评测有助于了解自然语言信息处理系统内部组成部分的性能，又叫作“内在评测”（intrinsic assessment）。

由于自然语言信息处理系统的语言文字评测基本上只涉及系统外在的总体性能，因此，主要采用黑箱评测的方法。

在评测时，对自然语言信息处理系统的输入和输出采取有区别的评测态度，采用“宽进严出”的策略。因为系统可能需要根据用户要求处理不规范的自然语言信息，系统的输入部分应允许存在不规范之处；系统的输出部分应严格规范。

语料库是为一个或多个应用目标而专门收集的、有一定结构的、有代表性的、可被计算机程序检索的、具有一定规模的语料的集合。它是按照一定的语言学原则，运用随机抽样方法，收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文库。从其本质上讲，语料库实际上是通过对自然语言运用的随机抽样，以一定大小的语言样本代表某一研究中所确定的语言运用总体（Feng 2006）。

语料库可以按照不同的方式而划分为不同的类型：

（1）按语料选取的时间划分，语料库可以分为历时语料库和共时语料库。

（2）按语料库的结构划分，语料库可以分为平衡结构语料库和自然随机结构的语料库。

（3）按语料库的用途划分，语料库可分为通用语料库和专用语料库。

（4）按语料库的表达形式划分，语料库可分为口语语料库和文本语料库。

（5）按语料库中语料的语种划分，语料库可分为单语种语料库和多语种语料库。

（6）按语料库的动态更新程度划分，语料库可以分为参考语料库（reference corpus）和监控语料库（monitor corpus）。参考语料库原则上不作动态更新，而监控语料库则需要不断地进行动态更新，以反映语言的动态变化和流通情况。

语料库可以从规范性、代表性、结构性、平衡性4个方面进行评测。

（1）语料库规范性的评测

语料库中的语料应该符合国家有关语言文字的规范[①]，如国家关于异体字、异形词、简化字、数字用法、标点符号用法、计量单位名称、异读词的规范。应当根据《第一批异体字整理表》《第一批异形词整理表》《部分计量单位名称统一用字表》《简化字总表》《GB/T 15834－1995 标点符号用法》《GB/T 15835－1995 出版物上数字用法的规定》《汉语拼音方案》《中国人名汉语拼音字母拼写法》《普通话异读词审音表》《中文书刊名称汉语拼音拼写法》等语言文字规范标准，设置相应的测试点，对语料库中的语言文字进行检查，以评测语料库符合规范的程度。

例如，可以设置如下的测试点来测试语料库中异体字、异形词的规范性：

异体字：汉语书面语中并存并用的同音、同义而书写形式不同的字，如："雇—僱"。应根据《第一批异体字整理表》测试其规范性，以"雇"为规范字。

异形词：汉语书面语中并存并用的同音、同义而书写形式不同的词语，如："按

① [1] 国际标准：ISO 7098: 1991 情报与文献—中文的罗马化[S] //国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 498-500.

[2] 国家标准：GB/T 12200. 1-90 汉语信息处理词汇 01部分：基本术语[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 240-251.

[3] 国家标准：GB/T 12200. 2-94 汉语信息处理词汇 02部分：汉语和汉字[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 252-275.

[4] 国家标准：GB3259-92 中文书刊名称汉语拼音拼写法[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 477-479.

[5] 国家标准：GB/T 15834-1995 标点符号用法[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 425-430.

[6] 国家标准：GB/T 15835-1995 出版物上数字用法的规定[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 431-435.

[7] 国家标准：GB/T 16159-1996 汉语拼音正词法基本规则[S]//国家语言文字规范和标准选编. 北京: 中国标准出版社, 1997: 491-497.

语—案语”“百废俱兴—百废具兴”，应根据《第一批异形词整理表》测试其规范性，以“按语”“百废俱兴”为规范词语。

对于语料库中不规范的语言现象，应当根据国家有关规范进行测试，以评测语料库规范的程度。但是，对于某些特殊用途的语料库，例如外国留学生汉语学习中介语语料库，就应以语料的真实性为主，而不强求对其进行规范。

（2）语料库代表性的评测

语料库对于其应用领域来说，要具有足够的代表性，这样，才能保证基于语料库得出的知识具有较强的普遍性和较高的完备性。

由于真实的语言应用材料是无限的，语料库的样本有限性这个特点是无法回避的。承认语料库样本的有限性，建设语料库时，在语料的选材上，就要尽量追求语料的代表性，要使有限的样本语料尽可能多地反映无限的真实语言现象的特征。语料库的代表性不仅要求语料库中的样本取自符合语言文字规范的真实语言材料，而且要求语料库中的样本要来源于正在“使用中”的语言材料，包括各种环境下的、规范的或非规范的语言应用。语料库的代表性还要求语料具有时代性，能反映语言的发展变化，能反映当代的语言生活规律。

只有通过具有代表性的语料库，自然语言处理技术才能让计算机了解真实的语言应用规律，才有可能让计算机不仅能够理解和处理规范的语言，而且还能够处理不规范的但被广泛接受的语言以及甚至包含有若干错误的语言。能否处理未经编辑或非受限的真实文本以及处理真实文本的数量，是衡量一个自然语言处理系统究竟是实用化系统还是实验性系统的试金石。

因此，语料库评测时，还应当指出语料库中存在的那些不规范但被广泛接受的语言现象以及包含有若干错误的语言现象，以反映语言文字使用的真实面貌。

（3）语料库结构性的评测

语料库是有目的地收集的语料的集合，不是任意语言材料的堆积，这就要求语料库具有一定的结构。

语料库必须是以电子文本形式存在的、计算机可读的语料集合。

语料库的逻辑结构设计要确定语料库由哪几个子库组成，要定义语料库中语料记录的码、元数据项、每个数据项的数据类型、数据宽度、取值范围、完整性约束等。

在语料库建设中，提倡采用通用的扩展标记语言XML（Extensible Markup

Language）来组织语料文件。采用XML语言组织语料库，可以减少程序和数据的依赖性，提高语料库的数据独立性，从而提高语料库的共享性。

使用XML语言组织语料库时，一个语料库的文件是一个或多个XML格式的文件集合，可以用DTD（Document Type Definition，文档类型定义）或者XML 模式（XML Schema）来定义它们的结构。这样，通用的软件（如IE5.0）就可以依据DTD来检查每个语料文件的结构是否规范，解读语料文件的程序就不用向传统的文件系统那样，过多地在程序中去解决物理存储结构的问题，从而提高语料数据和程序的独立性，提高共享性。

语料文件的形式可以是纯文本文件、XML格式的文本文件、关系数据库文件等，以便用户既可以利用语料库管理系统已提供的功能研究语料库，也可以在自己熟悉的软件环境下使用语料库。

（4）语料库平衡性的评测

在平衡语料库中，语料库为了达到平衡，首先要确定语料的分类指标，即平衡因子。平衡因子是影响语料库代表性的关键特征。

影响语言应用的因素很多，例如，语体，年代，文体，学科，登载语料的媒体、使用者的年龄、性别、文化背景、阅历，语料的用途（公函、私信、广告）等。不能把这所有的特征都作为平衡因子，只能根据实际需要来选取其中的一个或者几个重要的指标作为平衡因子。最常用的平衡因子有学科、时间、文体、地域等。应该根据平衡语料库的用途来评测语料库所选择的平衡因子是否恰当。

随着计算机技术的发展，语料库的规模正在变得越来越大。大规模的语料库对于语言研究，特别是自然语言处理研究具有不可替代的作用。但是，随着语料库的增大，垃圾语料带来的统计垃圾问题也越来越严重。而且，当语料库达到一定的规模后，语料库的功能并不会随着其规模同步增长。因此，应当根据实际的需要来评测语料库的规模，语料库规模的大小应当以是否能够满足其需要来决定。

语料的元数据可以反映语料库的基本信息。

元数据（Metadata）可泛义地理解为关于数据的数据或关于数据的信息。语料的元数据对于语料库语言学研究具有重要的意义，可以通过元数据了解语料的时间信息、地域信息、作者信息、文体信息等各种相关信息；也可以通过元数据形成不同的子语料库，满足不同兴趣的研究者的个性化需要；还可以通过元数据对不同的子语料库进行比较，研究和发现一些对语言应用和语言发展可能产生影响的因素；元数据还可以记录语料的知识版权信息，记录语料库的加工信息和管理信息。

语料库系统的评测

语料库元数据的评测应当遵循如下原则：

（1）简单明了、面向用户

一般来说，语料库用户不可能花很多精力去学习和掌握复杂的标注格式，因此，语料库的元数据要尽量接近日常的语言习惯。

（2）有弹性

语料库的篇头标注信息除了语料的知识版权信息、语料创建者的背景信息、语料载体的发行信息、语料的内容信息、语料的采样方式信息（书面语料或者口头语料）、语料的管理信息等共同项外，不同的语料库还有其各自特殊的要求，语料库的元数据的标准需要定义共同的数据项、命名规则、数据类型、数据宽度。在具体标注时，设计人员可以选择其中的一些项目，这些项目要遵守规范的约定，设计人员另外还可再增加一些别的项目。

（3）用标准的英文单词定义元数据项名

元数据项命名时，最好用西文符号，因为有些软件在解读数据时不支持中文的变量名。另外，为了国际交流的方便，应尽量用标准的英文单词定义元数据项名，而不要使用汉语拼音的简写。

（4）机器可读

标注后的语料库的元数据要能被通用的计算机程序解读，而不应另行专门编写程序来解读，这是实现语料库可共享、可集成的关键。用目前流行的文本标记语言XML来标注语料，可以部分达到这个目标。

（5）遵守元数据定义的国际标准

语料库元数据规范的制定应该遵守元数据定义的国际标准，并以之作为共同的规范标准。

提倡使用国际通用的XML语言来组织语料库中的语料。

语料库的自动切词和自动标注是语料库自动处理的一项重要内容。

语料库的自动切词就是使用计算机把连续汉字文本中的单词切出来，使单词与单词之间出现空白。语料库的自动标注就是使用计算机给切分后的各个单元标注上正确的词类和其他语法、语义信息。

语料库自动切词和自动标注的评测应当遵守如下原则：

（1）语料库的自动切词的评测应当遵循国家标准《GB/T 13715–1992 信息处理用现代汉语分词规范》（以下简称《分词规范》）。

（2）对于具体词语的切分，在考虑规范仍然举棋不定的情况下，可以参照《现代汉语词典》来决定。例如，“立功/的/机会/有的是/。”中的“有的是”在《分词规范》中没有规定，但是在《现代汉语词典》中收录了，就可以将“有的是”作为一个切词单位。

（3）不同的应用对象对于切词的颗粒度的要求不完全相同，为了兼容不同词语的颗粒度，可以容许同一语言结构可以按照不同的层次切分。例如，“工具箱”可以切分为“‘工具’+‘箱’”，也可以算为一个切分单位“工具箱”，这时，可以使用多层次的括号式表示为[工具/n箱/n]n（其中n为名词）。

（4）应注意切分时的歧义。切分歧义主要表现为：

交集型歧义切分字段：例如，“从小学”在“从小学电脑”中应当切分为“从小/学”，在“从小学毕业后”中应当切分为“从/小学”。

多义组合型歧义切分字段：例如，“将来”在“他将来北京工作”中应当分别切分为“将”和“来”，在“情况将来会改变”中不能切分，应当为一个单词“将来”。

（5）应注意命名实体（人名、地名、机构名）的正确切分。由于大多数的命名实体都不会存储在机器词典中，切分时容易出现错误，应当把命名实体的切分作为语料库自动切词评测的重要内容。

（6）标注时应当注意区分兼类词，选择正确的词性标注[①]。

例如：路很直（“直”应当标注为a）

他直哭（“直”应当标注为d）

又如：我在家（“在”应当标注为v）

我在办公室开会（“在”应当标注为p）

再如：他从日本回来（“从”应当标注为p）

我从不抽烟（“从”应当标注为d）

---

① 词类标记说明：n—名词，nh—人名，u—助词，vl—系动词，w—标点符号，d—副词，c—连接词，nd—方位词，p—介词，a—形容词，r—代词，k—后加成分，vu—助动词，vd—趋向动词。

（7）根据上述原则，采用手工或者半自动的方法制定评测语料的标准答案，作为标准切词和标准标注。

下面是国家语委现代汉语语料库切词和标注语料的样例，可以作为评测时的标准答案：

鸟/n的/u世界/n

杨栋/nh

鸟/n是/vl[大/a自然/n]n的/u歌手/n，/w鸟语/n[就/d是/vl]vl[大/a自然/n]n的/u音乐/n和/c诗歌/n了/u。/w

山村/n里/nd的/u鸟/n除了/p麻雀/n，/w就/d数/v燕子/n多/a了/u。/w[村/n人/n]n对/p燕子/n很/d爱护/v，/w说/v它/r吃/v庄稼/n的/u害虫/n，/w常/d吓唬/v[孩子/n们/k]n不要/vu去/v玩/v燕子/n，/w会/vu坏/v自己/r的/u眼睛/n。/w有时/d光/a屁股/n的/u小/a燕/n掉/v下来/vd，/w也/d要/vu送回/v[燕/n窝/n]n里/nd去/vd。/w

可以看出，在上面的标注中，“[大/a自然/n]n，[就/d是/vl]vl，[村/n人/n]n，[孩子/n们/k]n，[燕/n窝/n]n”等词语的标注，就采用了颗粒度不同的标注结果。

语料库的自动切词和自动标注的评测在很大程度上依赖于词典，由于参测语料库系统的词典中的词条和词的颗粒度不完全相同，因此，有必要明确地给出有关定义。

（1）正确切词和错误切词

如果切词序列$S_iS_{i+1}...S_j$中的汉字序列与切词序列$S_{i1}S_{i1+1}...S_{j1}$中的汉字序列一致，则称切词序列$S_iS_{i+1}...S_j$与切词序列$S_{i1}S_{i1+1}...S_{j1}$相等，记为$S_iS_{i+1}...S_j = S_{i1}S_{i1+1}...S_{j1}$。例如，工具/n箱/n = 工具箱/n。

在给定的语料中，设其字符串的基本汉字序列为$W_1W_2...W_n$，如果在标准答案中存在切词序列$S_iS_{i+1}...S_{i+k}$，使得$X = S_iS_{i+1}...S_{i+k}$，则称$X = W_1W_2...W_n$为被测语料库系统的一个正确切词。这时，X的正确切词数为k+1。

例如，“工具/n箱/n”是一个正确切词，正确切词数为2；“工具箱/n”也是一个正确切词，正确切词数为1。

不正确的切词，称为错误切词。

（2）正确标注和错误标注

对于被测语料库系统的正确切词X，X的词性标注是T，标准切词$S_{i1}, S_{i1+1}, ..., S_{i1+k}$的词性标注依次是$T_0, T_1, ..., T_k$，若T与$T_0, T_1, ..., T_k$一一匹配，则X的标注为正确标注，且正确标注数为k+1，由于正确标注之前必须正确切词，所以，这个k+1也可以称为正确标注切词数；若T与$T_0, T_1, ..., T_k$不完全匹配，则不匹配的标注为错误标注，不匹配的标注数就是错误标注数。在通常情况下，切词的颗粒度越大，其词性标注的歧义越小。当存在兼类词的情况下，应该进行兼类词判断，从若干个标注中选择出一个正确的标注。

我们建议使用如下公式来计算切词和标注的结果：

（1）切词正确率（segment right rate），用SR表示，公式为：

$$SR = \frac{SRN}{Sum}$$

其中，SRN是被测语料中正确切词数，Sum是与被测语料对比的标准切词中切词的总数。

（2）词性标注正确率（tagging right rate），用TR表示，公式为：

$$TR = \frac{TRN}{Sum}$$

其中，TRN是被测语料中正确标注切词数，Sum是与被测语料对比的标准切词中切词的总数。由于切词错误时标注就没有意义了，TRN是指被测语料中切词和标注都正确的数目，所以，总是有TRN≤SRN。

（3）词性标注相对正确率（relative tagging rate），用RR表示，公式为：

$$RR = \frac{TRN}{SRN}$$

其中，TRN是被测语料中正确标注切词数，SRN是标准切词中正确切词数。

基于同一标准语料测试的结果具有可比性。在上述定义中，Sum是标准语料的切词总数，SRN和TRN都是参照标准语料的切词得到的，因此，被测试语料的正确切词越多，SRN就越大，SR也就越大。同理，TRN、TR、RR也是如此。

例如，被测语料的切词和标注结果如下：

鸟/n的/u世界/n

杨栋/nh

鸟/n是/vl大自然/n的/u歌手/n，/w鸟语/n就/d是/vl大自然/n的/u音乐/n和/c诗

歌/n了/u。/w

山村/n里/nd的/u鸟/n除了/p麻雀/n，/w就/d数/v燕子/n多/a了/u。/w村人/n对/p燕子/n很/d爱护/v，/w说/v它/r吃/v庄稼/n的/u害虫/n，/w常/d吓唬/v孩子/n们/k不要/vu去/v玩/v燕子/n，/w会/vu坏/v自己/r的/u眼睛/n。/w有时/d光/n屁股/n的/u小/a燕/n掉/v下来/vd，/w也/d要/vu送回/v燕/n窝里/n去/vd。/w

与标准答案相比，标准答案中的“大自然”切为“[大/a自然/n]n”，被测语料切为“大自然/n”，是正确的切分；标准答案中的“就是”切为“[就/d是/v1]v1”，被测语料切为“就/d是/v1”，是正确切分；标准答案中的“村人”切为“[村/n人/n]n”，被测语料切为“村人/n”，是正确切分；标准答案中的“孩子们”切为“[孩子/n们/k]n”，被测语料切为“孩子/n们/k”，是正确切分；标准答案中的“燕窝里”切为“燕窝/n里/nd”，被测语料库切为“燕/n窝里/n”，是错误切分，因为这样的错误切分涉及两个单词，算为两个错误切分。在被测语料中总共有77个切分单位，故Sum=77，有两个错误切分，故正确切词数为75，SRN=75，所以，

$$SR = \frac{SRN}{Sum} = \frac{75}{77} = 0.9740 = 97.40\%$$

与标准答案相比，标准答案中的“光”是名词和形容词兼类词，标准答案标注为a，而被测语料标注为n，是错误标注；如果存在切词错误，标注也就没有价值了，所以，切词错误数也应当被包含在标注错误数之内；在这种情况下，被测语料的错误标注数为3，正确标注切词数为74，故TRN=74，所以，

$$TR = \frac{TRN}{Sum} = \frac{74}{77} = 0.9610 = 96.10\%$$

与标准答案相比，TRN=74，SRN=75，所以，

$$RR = \frac{TRN}{SRN} = \frac{74}{75} = 0.9866 = 98.66\%$$

本文中提出的这些意见都是推荐性的，仅供大家参考。目前我国语料库系统开发的研究很热烈，可是对于评测问题的研究还比较少，我们希望大家注意评测问题的研究。