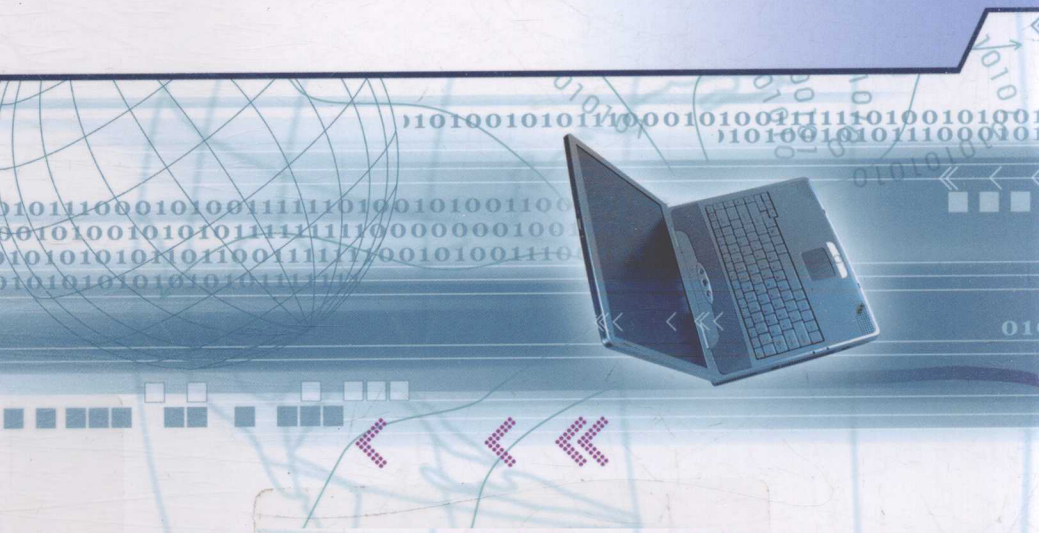
中国共产党思想理论资源数据库

人民金典语义查询系统

偏正属类语义标注规则



沈水荣　编著

人民出版社

中国共产党思想理论资源数据库

人民金典语义查询系统

偏正属类语义标注规则

沈水荣　编著

人民出版社

责任编辑:吴学金
装帧设计:涂　潇

图书在版编目(CIP)数据

人民金典语义查询系统
偏正属类语义标注规则/沈水荣 编著. -北京:人民出版社,2011.1
ISBN 978-7-01-009588-2

Ⅰ.①偏… Ⅱ.①沈… Ⅲ.①中国共产党-思想建设-数据库系统-情报检索 Ⅳ.①D261②G252.7

中国版本图书馆 CIP 数据核字(2011)第 000537 号

人民金典语义查询系统

偏正属类语义标注规则

RENMIN JINDIAN YUYI CHAXUN XITONG
PIANZHENG SHULEI YUYI BIAOZHU GUIZE

沈水荣　编著

人民出版社 出版发行
(100706　北京朝阳门内大街 166 号)

北京瑞古冠中印刷厂印刷　新华书店经销

2011 年 1 月第 1 版　2011 年 1 月北京第 1 次印刷
开本:880 毫米×1230 毫米 1/32　印张:1.625
字数:35 千字　印数:0,001-1,000 册

ISBN 978-7-01-009588-2　定价:5.50 元

邮购地址 100706　北京朝阳门内大街 166 号
人民东方图书销售中心　电话 (010)65250042　65289539

前　言

《偏正属类语义标注规则》是在人民出版社承建“中国共产党思想理论资源数据库”的过程中，专门为自主研发的“人民金典语义查询”系统所编著的。

三年前，在工程建设启动的时候，我们就动议开发有关软件系统，实现党和国家重要文献的语义查询。经调查发现，国内外对于计算机自动标引和语义自动识别的研究者众多，并提出过种种解决方案，但语义检索的质量和效果不够理想。目前计算机检索主要还是关键词检索，要达到完全意义上的知识点检索即语义自动识别，路程还很遥远。为此我们设想，在单纯依靠科技的力量还达不到理想效果的情况下，采用适当介入人工的办法不失为一个好的解决方案。

为此，我们研发了一个标注软件，开始请有关专家教授和研究生对领袖著作进行逐段逐句的标注，可以说这是一种“半自动标引”。起初对这项工作想得比较简单，过程中发现这里面道道很深，要解决的细节问题很多，比如，标注方法及用语如何与用户的查询习惯相吻合，如何按文章本身的内在逻辑划分知识点，如何把实现用户需求与计算机的实现能力结合起来，等等。标注人员提出的问题层出不穷。因此，头两年的标注同时也是一个形成规则的过程。标注规则从开始的几条简单约定，到逐步积累成现在这个小册子。《规则》中的每一条规定，每一个用词，都是从标注实

践中提炼出来的。所以,《规则》是实践和创新的结果,是“人民金典语义查询”系统不可缺少的一部分。

早在2009年5月,依据这套《规则》标注而成的“人民金典语义查询”系统已在“人民出版社网”上线运行,并向前来视察的李长春、刘云山、柳斌杰等领导同志进行了演示。目前已实现了对《邓小平文选》、《江泽民文选》和胡锦涛同志一系列重要讲话的单行本近10万多个知识点的语义检索。该系统同时以电子出版物的形式出版发行,被评为第三届“中华优秀出版物(电子出版物)奖”。它目前已拥有常用用户5000多人。这证明该查询系统及标注规则是可行的。

《规则》以语法逻辑中“偏正”、“属类”规则的运用为基础和核心,深入研究了人的思维规律、语言规则和查询习惯,在与软件系统研发的互动中编写而成,是一个严密而科学的体系。它是做好“人民金典语义查询”系统标注工作的基本依据,并用于对标注人员进行培训。希望标注人员务必在认真阅读、熟练掌握的基础上从事标注工作。

由于采用这种方法进行语义标注是一项全新的工作,《规则》还有不尽完善之处,希望大家在标注实践中认真总结经验,提出宝贵意见和建议,我们将不断加以修改完善。

采用这套标注法进行语义标注,其意义不仅仅在于取得直接的标注成果,还在于其标注成果可以为计算机自动标引提供语料库,用来建模和测试,使计算机通过“学习训练”不断提高自动识别的准确率。在全社会普遍推广这种做法,可能是实现完全意义上计算机自动识别语义的必经阶段。

这套《规则》的编写,得到了黄翬同志的密切配合和王晓峰、吴学金等同志的大力支持。中央国家机关有关部门和中央党校、

清华大学、北京大学、中国社会科学院，以及人民出版社内部共100多位专家教授、党政干部、青年学生也热心参与了讨论，提出了许多有价值的意见。在此特向大家表示衷心感谢！

编　著　者

二〇一一年一月

目　录

一、总　则

深刻领会**文章内容**，把握**用户需求**和**查询习惯**，了解本系统**检索规则**，是标注工作的根本要求。

领会文章内容，就是要对文章进行反复研读，并联系历史背景、讲话对象等，深刻理解、准确反映从文章整体到每个部分、每句话所含的语义以及相互之间的关联性。

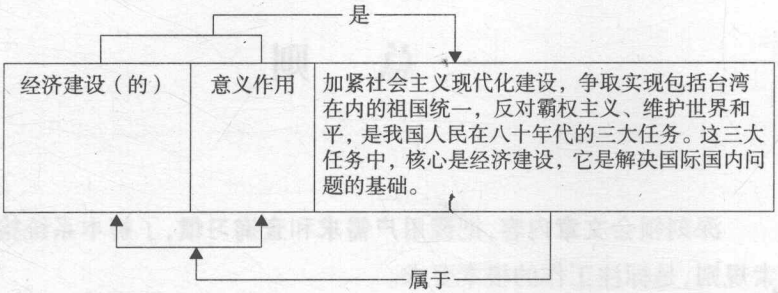
把握用户需求和查询习惯，就是要求标注者进行换位思考，了解文章中哪些知识点用户会查，用户可能输入什么样的关键词去查。做到用户怎么查，标注点就怎么做。

了解检索规则，就是要弄清楚本软件系统设置的一整套计算机检索规则，了解本规则设置的依据，从而自觉严格地按规则进行标注。

在此基础上，标注工作必须遵循以下三条总的原则：

（一）**匹配关系**。“**偏正**”、“**属类**”是匹配范围词、定点词和内容三者之间关系的基本规则。（1）**范围词与定点词**之间必须是偏正关系。即：范围词后面要能够通过一个“**的**”字，与定点词（所包含的事物性质状态及表达形式）匹配起来，反映所标注内容的意思。当范围词与定点词搭配比较勉强时，在范围词前面假设加上“关于”二字能匹配起来，也可作为正确标注。（2）所标注的**内容**与“范围词+定点词”之间必须是属类关系。即：**标注内容**必须属于“范围词+定点词”所指语义，也就是标注内容要能够通过“**属**

于”二字,与“范围词+定点词”搭配起来。如:



需要注意的是,一个标注点中有多个范围词条和多个定点词时,每个词条都要能够分别通过**每一个**定点词与所标注内容进行匹配。

(**二**)**两词分工**。范围词和定点词各有各的职能。范围词表示某一特定**事物本身**;定点词表示这一事物的**性质状态**或**表达形式**。

举例:

下面这段话是讲党的十二大历史背景,其中“党的十二大”是事物本身,可作范围词;“历史背景”是事物的性质状态,可作定点词(标“依从由来”)。

党的十二大	依从由来	现在这次代表大会和八大时的情况有了很大的不同。正如七大以前,民主革命二十多年的曲折发展,教育全党掌握了我国民主革命的规律一样,八大以后社会主义革命和建设二十多年的曲折发展也深刻地教育了全党。……

范围词的标注必须给定点词留有标注空间。范围词的功能只是框定语义所在的范围,**定点词**才能最后完成整体语义标注。试

图用范围词把一条内容的语义说完整、说清楚，势必造成范围词、定点词之间的**重复标注**。

如，下列标注点把表示事物性状的词（历史背景）标进了范围词，它与定点词“依从由来”进行匹配，等于是说“历史背景”的“历史背景”，发生了同义反复，会使用户查询时发生混乱。

党的十二大的历史背景	依从由来	现在这次代表大会和八大时的情况有了很大的不同。正如七大以前，民主革命二十多年的曲折发展，教育全党掌握了我国民主革命的规律一样，八大以后社会主义革命和建设二十多年的曲折发展也深刻地教育了全党。……

如下列两个标注点犯了同样的错误：

党的第七次全国代表大会的历史地位	意义作用	一九四五年在毛泽东同志主持下召开的党的第七次全国代表大会，是建党以后民主革命时期我们党最重要的一次代表大会。……那次代表大会，为新民主主义革命在全国的胜利奠定了基础。

新民主主义革命胜利的基础	依从由来	一九四五年在毛泽东同志主持下召开的党的第七次全国代表大会，是建党以后民主革命时期我们党最重要的一次代表大会。……那次代表大会，为新民主主义革命在全国的胜利奠定了基础。

（三）**有限标注**。每一篇文章的语义信息都会十分丰富，我们很难也没有必要把所有语义都标注出来。标注工作必须在**有限范围内**进行，这种有限性具体表现在：一是规定的可作**范围词**的词组类型是有限的，并非任何词语都可以入范围词；二是**定点词**是有限

的，必须在固定的词组内选择；三是**范围词、定点词、标注内容**之间的关系是确定的，受到“偏正”、“属类”匹配原则的限制。受到上述三方面限制，就有一些知识点是无法标注出来的，因此可以**不标**。同时，对一些虽然按规则可以标注，但在文中不是**直接表达出来、不是中心语义、查询价值不大**的内容，也可以不标。本查询系统设置“全文检索”功能，就是为了向用户提供无法标注的一些内容的查询。

二、标注点的划分

标注点划分的基本要求：**一是循序标注**。对一篇文章，应按照从**前**到**后**、从**总**到**分**、从一段文字中所表达的**基本意思**到这段文字中所包含的**各个具体知识点**的顺序，依次、逐一地进行标注，以体现出整篇文章标注的层次和节奏。具体说，首先要从**整体上**做好标注；然后对每**一个段落**或每**一个层次**进行标注；最后对**语句**进行标注。但是，当文章的某一段落、层次不能用合适的范围词和定点词进行标注时，可不作为一个整体来标注。标注要力求全面，经典著作、重要文献中**绝大多数文字**都有必要进行标注，防止出现**重要观点**和**重要内容**的漏标。**二是标注中心语义**。标注一段话，一定要抓住这段话的**总体语义**、**中心语义**、**中心论点**，不要以论证过程中所用的**大前提**、**小前提**、**论据**作为标注点，以此代替总体语义、中心语义、中心论点；如果论证过程中的某些话需要单独作为知识点来标，那么要看这些知识点是否**明显**，含义是否**充实**，是否有**查询价值**。**三是尊重文章原意**。要按照文章原有的逻辑结构、脉络以及各知识点的内在关联切分好每一个标注点，保持每个标注点语义的完整性，准确反映出每个语段标注内容本身所具有的含义。**四是必须使用范围词**、**定点词切分知识点**。语义标注与划分文章的段落大意不完全一样。一篇文章的段落大意可以由读者**自己组织语言**来概括，而语义标注其范围词必须尽可能采用**文中的原话**，定点词必须在**规定的词目**中选择；文章的段落大意是用**句子**来概

括的，而语义标注的知识点一般是用**词或词组**来概括的。这样，一段文字只有当它能够选用恰当**范围词**和**定点词**作标注时，才能单独切分成一个知识点；反之，即使有明确的段落大意的一部分文字，也不能作为一个**标注点**。换句话说，一段话能不能作为一个知识点来切分，是由能不能选用**合适的范围词和定点词**来标注它决定的。

标注点划分有以下几种基本方法：

（一）**一语一标**。一段文字做成一个标注点，并在一个范围词下，只选取一个定点词进行标注。简要地讲，就是一条标注只用一个**范围词**和一个**定点词**。

举例：

现代化建设的目标	本质要义	为把我国建设成为现代化的，高度文明、高度民主的社会主义国家，为反对霸权主义，维护世界和平，推进人类进步事业，而努力奋斗。

（二）**一语多标**。一段文字内容有多项语义并存，可以从多个侧面理解其含义，这种情况下，可用两个以上的**定点词**或两个以上**范围词**进行标注。这有三种情况：一是一个范围词对**多个定点词**；二是多个范围词对**一个定点词**；三是多个范围词对**多个定点词**。

例1：

八十年代的三大任务	概念内容 本质要义 特定表述	八十年代是我们党和国家历史发展上的重要年代。加紧社会主义现代化建设，争取实现包括台湾在内的祖国统一，反对霸权主义、维护世界和平，是我国人民在八十年代的三大任务。这三大任务中，核心是经济建设，它是解决国际国内问题的基础。

例2：

市场经济规律/宏观调控体系/市场在资源配置中的基础性作用	意见要求	要深化对社会主义市场经济规律的认识，从制度上更好发挥市场在资源配置中的基础性作用，形成有利于科学发展的宏观调控体系。

例3：

转变发展方式/实现人均国内生产总值翻两番	意见要求 预见展望	转变发展方式取得重大进展，在优化结构、提高效益、降低消耗、保护环境的基础上，实现人均国内生产总值到二〇二〇年比二〇〇〇年翻两番。

（三）**同语另标**。当同一段文字含有**多种**语义，且不能用**相同范围词**或**定点词**标注时，另起条目进行标注。

举例：

转变发展方式/实现人均国内生产总值翻两番	意见要求 预见展望	转变发展方式取得重大进展，在优化结构、提高效益、降低消耗、保护环境的基础上，实现人均国内生产总值到二〇二〇年比二〇〇〇年翻两番。

人均国内生产总值翻两番	时地数序 依从由来	转变发展方式取得重大进展，在优化结构、提高效益、降低消耗、保护环境的基础上，实现人均国内生产总值到二〇二〇年比二〇〇〇年翻两番。

（四）**集群标注**。当一段文字包含的若干个内容**相近**，而且文句位置**互相挨着**的知识点，可以用相同定点词标注，但难以选定**相同范围词**进行标注时，可做**集群标注**，即：在同一条目中用**多个范**

围词标注相互并列的知识点。集群标注条目中的内容和范围词文字都不宜**过长**。

举例：

市场经济规律/宏观调控体系/市场在资源配置中的基础性作用	意见要求	要深化对社会主义市场经济规律的认识，从制度上更好发挥市场在资源配置中的基础性作用，形成有利于科学发展的宏观调控体系。

（五）**合并标注**。当同一段文字可以从两个以上侧面理解其含义，需要用不同范围词，但又可以用**相同**定点词进行标注时，可做**合并标注**。这也就是把范围词不同、标注内容以及定点词相同的标注点合并到一起进行标注。

举例：

广东/经济特区/深圳/珠海	状态情况；	一九八四年我来过广东。当时，农村改革搞了几年，城市改革刚开始，经济特区才起步。八年过去了，这次来看，深圳、珠海特区和其他一些地方，发展得这么快，我没有想到。看了以后，信心增加了。

要仔细辨别“集群标注”与“合并标注”的区别，“集群标注”的标注点内容是多个部分文字组成多个部分含义；“合并标注”的标注点内容是同一部分文字包含多个方面含义。

（六）**多处合标**。一篇文章中，范围词和定点词都相同的文字在**相近**的地方**两**处以上出现，且单独标注显得**零碎**时，可合并成一个标注点标注，中间用**省略**号。如合并标注的内容在不同的自然段中，中间须**换行**。