**Research On The Development Of Tagging Corpus**

**标注语料库的开发研究**

**By**

**罗晓雨**

**Supervised by**

**Lecturer Meng Sheng**

**January 10th, 2020**

**College of Foreign Languages and Cultures**

**Chengdu University of Technology**

**标注语料库的开发研究**

# 摘要

标注语料库系统具有一定的综合性，涉及到语料库用户、语料采集、语料库管理等各方面，而且各部分之间有着密切的联系，因此，在对标注语料库进行开发时，需要选择合适的技术方法。语料库的标注需要遵循一定的原则，包括独立性原则、公开性原则等，确保符合实际的标注需要，这样才能更好发挥出语料库标注的作用。基于这些情况，本文对标注语料库的开发进行了研究。

本文对语料库做出了简述，对语料库标记做出了分析，主要涉及到语料库标记的归责、模式等。对于标注语料库的开发，本文阐述相应的开发流程，从预处理、语义标注、语料入库方面进行了分析，以期为其他标注语料库的开发提供借鉴。

**关键词：**标注；语料库；开发

# 

# Abstract

The tagging corpus system has a certain degree of comprehensiveness, involving corpus users, corpus collection, corpus management and other aspects, and each part has a close relationship. Therefore, in the development of the tagging corpus, we need to choose appropriate technical methods. The tagging of corpus needs to follow certain principles, including the principle of independence, the principle of openness and so on, to ensure that it meets the actual needs of tagging, so as to better play the role of corpus tagging. Based on these situations, this paper studies the development of tagging corpus.

This paper gives a brief introduction to the corpus and an analysis of the corpus markers, mainly involving the attribution and mode of the corpus markers. For the development of annotation corpus, this paper describes the corresponding development process, and analyzes the preprocessing, semantic annotation and corpus storage, in order to provide reference for the development of other annotation corpus.

**Keywords:** tagging; corpus; development

# 目录

[摘要 I](#_Toc29028673)

[Abstract 2](#_Toc29028674)

[1 语料库简述 4](#_Toc29028676)

[2 语料库标记分析 4](#_Toc29028677)

[2.1 语料库标记的规则 4](#_Toc29028678)

[2.2 语料库标记的模式 4](#_Toc29028679)

[3 标注语料库的开发 5](#_Toc29028680)

[3.1 总体开发流程 5](#_Toc29028681)

[3.2 具体开发分析 6](#_Toc29028682)

[3.2.1 预处理 6](#_Toc29028683)

[3.2.2 语义标注 7](#_Toc29028684)

[3.2.3 语料入库 9](#_Toc29028685)

[结论 1](#_Toc29028686)0

[参考文献 1](#_Toc29028687)1

# 1 语料库简述

语料库是文本的有序集合，是各种分类，检索，综合和比较的基础。它是根据预定规则准备，编码或存储的机器可读文本或部分文本的有组织的集合。用于进一步语言分析或术语工作的原材料。经过科学选择和标注，适当大小的语料库可以反映和记录该语言的实际用法[1]。人们使用语料库对大型自然语言进行调查和计数，建立统计语言模型，研究和应用基于统计的语言处理技术，并在信息检索，文本分类，文本过滤和信息提取等应用中取得了进展。 语料库已成为语言理论研究，应用研究和语言工程必不可少的基础资源。

# 2 语料库标记分析

## 2.1 语料库标记的规则

（1）公开性原则

目前，还没有一个统一的标签原则和标签规范，每个人都可以接受。 因此，成熟的语料库必须具有详细的标签规范文件，包括标签原理，符号系统和语法规则。语料库标记的过程是理解和解释语料库的过程[2]。显然，只有使用详细的标注规范文件，用户才能阅读和理解带标注的语料库，并指导机器识别和理解带标注的语料库[3]。

（2）通用性原则

建立一个大型的标记语料库是一个巨大的项目。只有实现语料库标注的普遍性，我们才能追求可共享性，最大化语料库的价值并减少重复劳动。例如，经过专家的多次讨论，并使用多个自动分词软件进行多轮分词，标记实验和使用多个自动分词软件，才能获得国家现代汉语语料库标注规范“信息处理的现代汉语词性标记集规范” [4]。它吸收了语言学家的研究成果，并考虑了各种分类系统。它是一套基于信息处理实践的现代汉语词性标签集规范。它提供了现代中文书面词性标签集的符号系统，使各种中文信息处理系统可以使用各方都可以接受的更加统一的词性标签系统来标记各自的系统。

## 2.2 语料库标记的模式

人们使用各种标记模式来标记语料库，其中包括TEl。标签模式被认为可以最好地反映语料库语言学家当前的趋势，他们正在努力为机器可读文本信息的编码建立更正式的国际标准。许多大型语料库，例如BNC语料库，都采用了TE1标记模式。 TEI标签模型是由三个学术团体联合开发的，包括计算语言学会（ACI），文学和语言计算协会以及计算机与人文学会。根据TEI标注模式，话语由标题和正文组成[5]。文本是指与文本相关的背景信息，即本文中提到的元数据。 TEl标记模式具有两种基本的标记方法：标签和实体引用。话语由许多语言单元组成。一个语篇是由许多语言单位构成的，这些语言单位可以是词、句子、段落，也可以是一个章节，甚至是整部书。一个语言单位的开始用起始标记<⋯⋯>，尖括号内为标注附码，语言单位的结束用结束标记叫⋯>，尖括号内为标注附码。实体参考是一套采用缩略附码对语篇内语言单位的语言特征进行详细标注的方法，这些缩略附码被称为特征标注系统。用“＆”，“、”，“；”等符号作为附码开头的特征标记。一个语篇的总体标注情况用文件类型描述DTD标出。DTD提供关于语篇所包含的语言单位、语言单位的组合形式、附码的意义等信息。TEI标注模式已经为诗歌、书信、戏剧等基本文体制定了一套标注的标准。

# 3 标注语料库的开发

## 3.1 总体开发流程

标注语料库的建设主要分为预处理、基本加工、语义标注及语料入库几个阶段。入库后，系统为语料提供了访问控制接口。具体流程如图3.1所示。

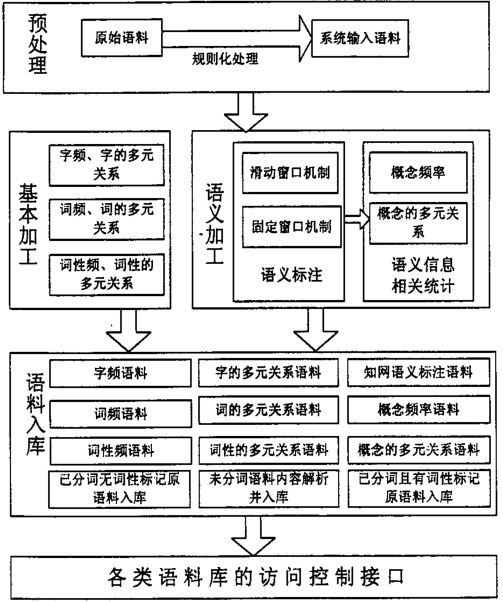


图3.1 标注语料库建设的总体流程

## 3.2 具体开发分析

### 3.2.1 预处理

语料库需要在进入系统之前进行预处理。 预处理主要负责规范源语料库。 所谓的正则化过程是根据自定义模板统一格式化已过度用词的原始语料库。在随后的工作中，将为带有模板形式的语料库提供一个通用的处理界面。模板形式如图3.2所示。

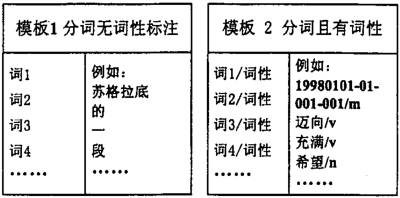


图3.2 语料模板格式

模板1被建议用于不带标记的语料库。每个分段的单词占用一行，而Lang使用回车和换行作为单词之间的分段标记。模板2被建议用于分段和词性语料库。每个单词及其词性占一行，单词及其词性之间用“ /”分隔。以这种方式自定义模板的原因主要来自两个方面。首先，在语义信息标记过程的第二阶段，我们需要在单词上添加语义标签。如果将空格或其他字符用作单词之间的分割符号，则很难向源语料库添加新的语义标记。我们使用回车和换行作为单词之间的分隔符，为语义标记留出足够的空间。其次，从结构化语料库文件的存储角度来看，语料库文件模板化过程实际上是对语料库内容进行结构化的过程，这为实现基于内容的语料库访问奠定了基础。在该系统中，不仅需要将源语料库规范化为统一的模板形式，而且从处理和处理中获得的所有新语料库都采用类似的模型，因此系统可以使用统一的方式访问所有语料库内容。

### 3.2.2 语义标注

分词后，将语义标记添加到语料库。对于带有词性标记的语料库，首先使用词性过滤器除去与词的词性不匹配的含义项，并简化语义关系挖掘。如果源语料库没有词性标注，它将直接进入语义关系挖掘机器的标注阶段。 经过处理后，系统将输出带有语义标记的语料库，但是某些单词仍然具有多义性。这时，可以使用概念频率和其他处理方法，准确地标注语料库。标记机器后，系统允许使用手动验证工具进行验证。标记的总体过程如下图3.3所示：

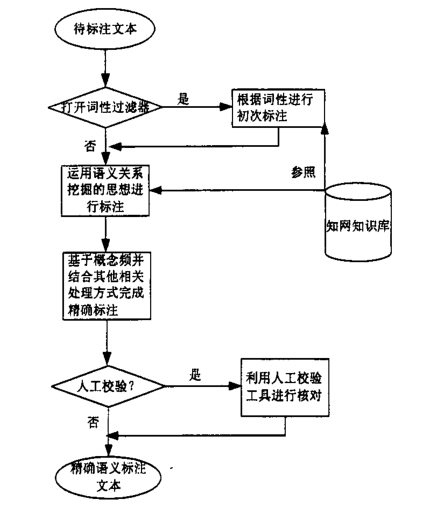


图3.3 语义标注流程

（1）词性过滤

由于词性标识与概念是紧密相关的，因此词性标注可以为多义词标注提供丰富的信息。在《知网》中如果不考虑词性约15％的实词为多义词，如果考虑词性的话约12％为多义词，也就是说其中约20％的歧义可以通过词性来消除，我们把这部分词称为伪多义词。在实际中，系统首先对多义词进行词性过滤，因此与多义词当前词性不相关的词义将不会进入系统，不会为多义词标注提供任何信息，即剔除伪多义词。对于词形存在于词典中而词性无法匹配的词在标注过程中系统保留该词的所有信息(词形、词性、词义等)。由于词性过滤可以将不相关的词义剔除掉，因此一旦词性标注错误，正确的词义将可能被从待选词义中过滤掉，从而造成标注错误。在这里，系统为词性过滤设了开关，当关闭词性过滤器时，所有的词义将不考虑词性全部进入系统。

（2）基于语义关系的标注

语义标注分为手动标注和自动标注。所谓的语义自动标记是指计算机使用逻辑运算和推理机制来正确确定和标记在特定上下文中出现的单词的含义。语义自动标记的研究是让计算机掌握语言知识并使用上下文信息确定单词的含义。使用语义关系挖掘进行自动语义标记的过程分为三个步骤，如图3.4所示。输入语料库是已被标记化的语料库，并参考知网知识词典用单词的所有相关含义标记进行标记。在语义关系挖掘阶段，充分挖掘每个词的语义，以获得所有含义的可能含义，形成含义序列，比较歧义词含义与其他含义的同现要确定上下文中单词的精确语义：对信息结构的分析，作为对原始含义的共现的补充，以进一步确定要区分歧义的单词的每种含义与上下文单词的相关性。最后，将以上两个方面的判断结合起来，消除了与上下文完全矛盾的含义项，达到了消除歧义的目的。因为某些单词可能仍然具有多义性，所以输出是部分歧义的语料库。

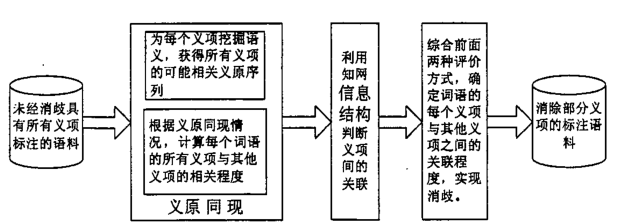


图3.4 语义标注流程图

### 3.2.3 语料入库

系统提供的语料库文件不是简单地整体存储在数据库服务器中，而是在将语料库文件存储在数据库中时，系统构造文件，自动解析文件内容并将其保存为文件属性。 Oro [e9iFS服务器基于文件内容属性实现访问。在Oracle9iE中，属性信息存储在不同的Oracle表中，而文档内容则以大对象的形式存储在表ODMM\_CONTENTSTORDE中的LOB类型中。如果使用系统默认值来存储文件，则语料库内容将作为一个非结构化的字节流作为一个整体存储，而解析文档内容的工作将由应用程序本身负责，因此基于文件的访问不能取得成就。但是Oracle9iF提供了解析器机制，该机制可以促进文件内容的分析。解析器是一个程序，可以将收到的消息字符串拆分为指定的组成元素。在iFS中，解析器是Java类。当文件存储在数据库中时，它可以自动从文件中提取结构化数据元素。同时，Orade9iFS允许用户基于现有文档属性类型进行扩展。

# 结论

本文主要对标注语料库的开发进行研究，对语料库、语料库标注做出了说明。在对标注语料库进行开发时，主要涉及到预处理、语义标注和语料入库方面，完成了相应的语义自动标注设计。标注语料库在多个领域均能发挥出较好作用，要想构建一个高质量的标注语料库，需要采取更为科学的方法、技术，这样才能更好发挥出标注语料库的作用。

语料库在各个领域中均有着较大作用，不同类型的语料库构建需要考虑到实际的构建目标，结合构建需求来进行标注语料库的开发，这样能为相关工作提供较大便利。

# 

# 参考文献

1. 王敬, 杨丽姣, 蒋宏飞, 等. 汉语二语教学领域词义标注语料库的研究及构建[J]. 中文信息学报, 2017, 31(1):221-229.
2. 刘玉安. 《语料库标注和分析中的计算机处理方法》介评[J]. 外语教育, 2017(1):149-153.
3. 齐芷玥. 语料库研究初探[J]. 文存阅刊, 2018(11):36-36.
4. 周惠巍, 杨欢, 徐俊利, 等. 中文模糊限制信息范围语料库的研究与构建[J]. 中文信息学报, 2017, 31(3).
5. 刘阳, 赵秀凤. 基于语料库工具Wmatrix的能源隐喻研究[J]. 现代语文(语言研究版), 2017(5):84-86.