中文信息处理开放平台的设计*

刘群[1,2] 张浩[1] 白硕[3]

[1](中国科学院计算技术研究所, 北京 100080)
[2](北京大学计算语言学研究所, 北京 100871)
[3](国家计算机与网络信息安全管理中心, 北京 100031);
E-mail: liuqun@ict.ac.cn

摘要: 我国的自然语言处理研究，在很大程度上处于一种低水平重复状态，由于缺乏一些公共的基础设施，很多研究工作都要花费大量的精力从底层模块做起，造成研究工作难以深入。本文提出，可以将开放式的开发模式应用于自然语言处理领域，并给出了一个面向中文的自然语言处理开放平台的设计。这个平台能够共享代码、语料、语言知识库等资源，并支持协作开发。这个平台的上层管理采用项目方式，实现了资源的重复利用。随着参与者的增多，和项目的发展，这个平台一定会为中文信息处理提供大量的资源。

关键词: 开放源码; 资源平台; 自然语言处理

引言

我国的自然语言处理研究，在很大程度上处于一种低水平重复状态，由于缺乏一些公共的基础设施，很多研究工作都要花费大量的精力从底层模块做起，造成研究工作难以深入。近些年来，随着 Linux 等开放源码软件的惊人发展，开放式开发的思想正在逐渐深入人心[1]。开放的好处不仅体现在成品上，更体现在过程中。只有当开发过程成为开放式的以后，该领域的工作者才能以最自然的方式形成最大规模的协作，朝着一个共同的目标努力，把一个个好的思路贡献出来，使得一个公共的产品迅速得到演化更新。

在自然语言处理领域，我们处理的都是同一个对象：人类语言。而且，这个对象体系庞大，从词法层次直到语义层次现象复杂。所以，要对这个公共的对象进行处理，理应存在一套公共的基础设施和工具集，理应存在最开放的协作和过程管理，否则，我们要进行大量的低水平重复开发，并且总是处在争执不下的局面，难以提高这个领域的处理水平。

目前，在自然语言处理领域，已经有了很多的共享资源，特别是英语的资源已相当丰富，词典、语料库、词法分析、句法分析、命名实体分析等很多基础性的研究领域都有了可共享的资源，这使得相关的研究工作起点很高，工作容易深入。不可否认，我国的自然语言处理领域，各种自然语言处理的基础资源建设也有了长足的进步。其中最有标志性意义的两个事件是董振东先生的《知网》和北京大学、人民日报社和富士通公司的《人民日报标注语料库》的公开发布，这对中文信息处理的研究起到了极大的推动作用。不过，与英语或者日语相比较，我们可以得到的可共享资源还是要少的多。这极大地妨碍了我国自然语言处理研究的进展。一个典型的现象就是，几乎所有从事相关研究工作的人都要自己开发一套分词系统，这就导致我国的分词研究低水平重复式地长盛不衰，而一些更加深入的研究工作，如句法分析、语义分析等等，却总是难以深入。

*本课题受国家重点基础研究项目（973）资助（G1998030510 和 G1998030507-4）

本文当中，我们提出采用类似 Linux 的开放源代码方式，建设一个自然语言处理的开放平台。这种方式的好处不仅仅在于开放和共享，我们认为一个更大的好处在于，可以吸引一批真正有志于此领域的研究工作者，大家通力协作，完成一些大家在孤立状态下难以完成的工作。

本文中我们将探讨建设一个面向中文的自然语言处理开放平台的若干问题，包括其目标、意义和组织结构、整体设计，以及平台之上的项目规划。

1. 开放资源

建设一个开放平台，首先要把开放的资源加以明确。我们把自然语言处理的开放资源定义为以下几类：

(1) 代码资源

目前各个领域都已有大量的开放源码计划。在中文信息处理领域方面，我们只在国外少数几个网站[6]找到了很少的中文处理源代码，其中最复杂的是一个用 Perl 语言编写的汉语词法分析器，具有初步的词语切分和人名识别功能，正确率不高。其他方面几乎都还是空白。

我们认为，自然语言处理的代码资源主要应包含以下几种类型：

a) 基础型：发展基础的和通用的技术。建立评测数据集和评测结果记录，包括功能指标，如精确率、召回率和性能指标，如运行时间、空间开销，吸引参与者改进源码或提交同类源码，促进目标系统的演化发展。例如，分词和标注系统[3,4]、短语划分和句法分析系统[5]。

b) 工具型：自然语言处理在底层有着许多琐碎的任务。推动工具共享，能够使我们尽快走出作坊式加工的现状。例如，词典管理工具，汉字编码处理工具。

c) 应用型。制定小型应用目标，通过其实用性来吸引更加广泛的参与者。这些小的应用都有可能发展为具有真正应用前景的系统，关键是要公开接受检验。例如，汉语的书写检查工具、拼音汉字转换、人机聊天系统、特定领域的知识问答系统。

(2) 语言资源

自然语言处理越来越依赖于语言资源的丰富性。目前网上已经有了一些可以共享的中文语言资源。除了前面提到的《知网》和《人民日报标注语料库》外，LDC（Linguistic Data Consortium）网站上也提供了很多中文的资源，其中有些也是免费的。

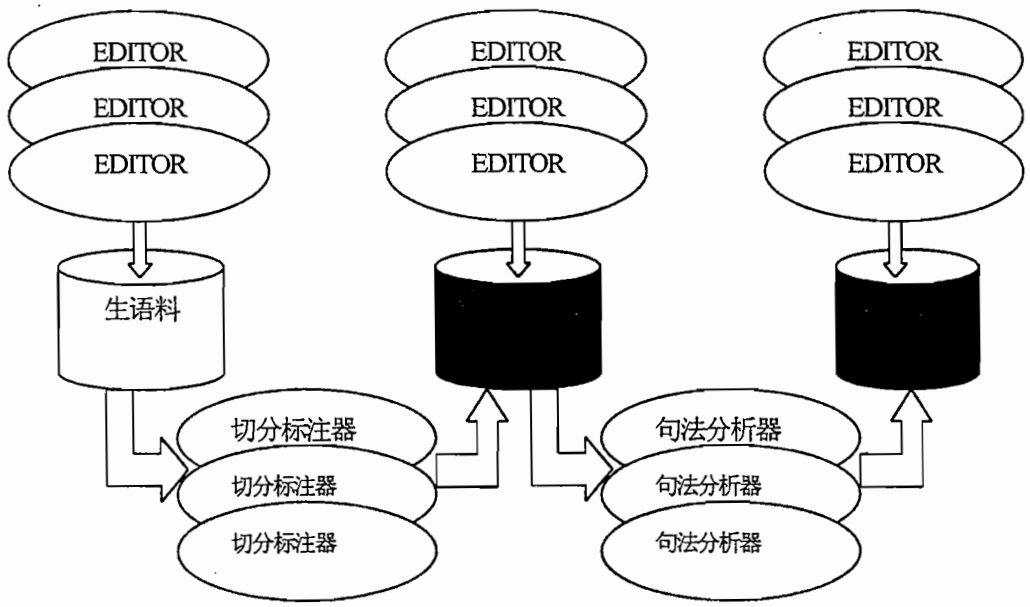

图 1 语料库资源的级联式加工模型

语言资源多种多样。对于中文的分词系统来说，所需的资源包括切分标注好的语料库、各类型的专名库，词典等；对于句法分析系统来说，除了与词法部分共享的词典资源外，还需要语法规则库、进行过句法标注的语料库——树库[2]。 尽管类型多样，我们还是可以初步地将语言资源分作语料资源和语言知识资源两种。语料资源主要是不同层次的语言现象的样本集。语言知识资源主要是语言知识库。

和开放源码的“Given enough eyeballs, all bugs are shallow”的思想[1]相平行，把语言资源放置于众人审视的目光之下，最有利于资源质量的提高，同时也最有利于规模的扩大。

以语料库资源的级联式加工模型为例，我们可以设想一个多机并行，人机互助的语料库加工过程，如图 1 所示。开放利于发展。语料资源如此，语言知识库资源也不例外。以语法规则库为例，就是需要很多人讨论一道来调整的知识库。可以说，使规则系统完善的最好方法：将其开放，经受检验。

(3) 文档资源

不论是自然语言处理的研究还是开发，都涉及到大量的关于理论、模型、系统的文档，这些文档的重要性不亚于前面任何一种资源。推动自然语言处理作为一门学科的发展，迫切地需要知识的积累。建立一个由参与者共同维护的文档库，必将发挥每个个体的知识优势，实现文档的“准”和“全”两大目标。实际上，Linux 下面的文档计划和源码计划同步开展，已经提供了成功的案例。

自然语言处理领域，文档大致可以分为以下几类：

a) 文献：各种公开发表的学术文章和技术报告。由于各方面条件的限制，国内的研究者，特别是广大的学生，很难比较全面地接触到国际上相关的研究资料，有时甚至一些重要的文献都没有掌握，这对从事相关的研究工作非常不利。收集、整理相关的各种文献资料，应该是自然语言开放平台的一项重要任务。对于国内研究者来说，翻译也是一项重要的任务。

b) 标准：国际国内在自然语言处理方面都出台了一些标准，如国内的信息处理用汉语分词标准、分词标记集标准等等；

c) 技术资料：一些大型系统开发中所形成的使用说明、技术规范、规格说明书等等，在系统开发者同意的情况下，可以放到开放平台上。这种资料是开发者实际工作的经验积累，对其他研究者有重要的借鉴意义。

对于文档类资源，开发平台应提供完善的管理和检索功能。

2. 开放平台

建设一个支持对以上这些资源实施开放的平台，需要提供哪些功能呢？

以上这些资源都可以抽象为文件或文件集合。需要编辑内容的资源主要都是纯文本，或可以转换为纯文本格式的。所以，并发式的文本编辑功能是最需要的。

如果这些资源没有好的组织，不仅混乱，而且发挥不了价值。一个凌驾于资源管理之上的项目管理功能十分必要。这个项目管理可以理解为资源的视图，或是组合。

最后就是关于使用者的管理——用户管理。用户管理是与项目管理相协调的。用户是项目的参与者，项目由用户来共同维护和开发。

3. 组织结构

我们的目标是建设一个支持资源共享和开发的平台，具有工作平台的性质。所以，我们认为，按照项目方式进行组织是比较合理的。我们希望吸引对此感兴趣的志愿者在平台上管理和组织自己的项目。

整个平台按照资源包和项目两个层次来组织。

平台上所有的资源组成一个个的资源包，这些资源包按照前面的介绍分为代码资源包、语言资源包和文档资源包三类。

每个资源包都唯一地归“属于”一个项目，我们称该项目“拥有”这个资源包。一个项目可以“拥有”一个到多个资源包，项目对于它所“拥有”的资源包有读写权限。同时一个项目可以“使用”一个到多个不“属于”它的资源包，项目对于它所“使用”的资源包有读权限。一般情况下，资源包都应该允许被不“拥有”它的其他项目所“使用”，从而达到资源充分共享的目的。

一个项目有唯一的一个项目管理员，可以有多个项目参与者。从项目管理员、项目参与者和普通访客的角度，可以看到不同的项目视图：

(1) 项目管理员视图：

可以看到项目的全部资源，以及各个用户对资源所进行的修改的提交。提供定版功能，管理员可以由此决定哪些修改应被保留下来将资源更新。项目管理员可以决定加入和删除该项目的参与者。

(2) 项目参与者视图:

可以看到项目的全部资源。提供提交/上载的接口。

(3) 项目访客视图:

只能看到权限允许的部分资源。

下图为项目和资源包关系的示意：

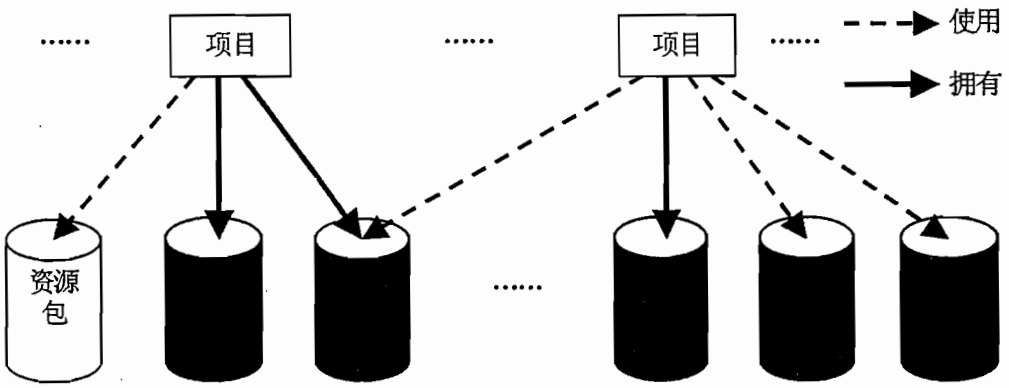

图2 项目和资源的关系

4. 整体设计

1. 前台目录

用户登录后进入相关项目，如果该项目具备相应资源，并且用户具备所需权限，则可能看到以下内容：

- 项目介绍
- 项目新闻：最近进展、动态
- 下一步计划：TODOLIST
- 代码资源

CVS：代码的并发编辑环境。

下载：打包下载新版本程序。

更新记录：对程序所进行的最新更新的报告。

- 语言资源

CVS：语言资源的并发加工环境。
下载：打包下载加工好的语言资源。
更新记录：对语言资源进行的最新更新的报告，包括详细的条目说明，便于引起关注。
- 文档资源

提交接口：给出文件及详细描述。
分类目录和关键词检索接口
- 演示:：本项目的在线演示版本

问题报告
- 论坛

2. 后台管理

开放平台建立在一个Linux服务器上，客户端可以使用Linux、Unix或Windows平台。开放平台上的项目运行平台可以由项目任意指定。

更为重要和困难的是资源和项目的管理如何实现。开放平台的管理主要通过一个数据库和一个源代码版本管理软件来实现。

数据库主要是用来进行用户管理、资源包管理和项目管理的。用户与项目的参与关系以及项目和资源的隶属关系是数据库中的主要关系。数据库目前采用MySQL。

源代码版本管理软件用于具体的资源文件的管理，可以实现资源文件的历史记录保存、版本比较、多人协同开放等等。我们目前使用CVS作为开放平台的源代码版本管理工具。作为一个源代码版本管理软件，CVS在以Linux为代表的开放源代码运动中起到了重要作用。与Microsoft的Visual SourceSafe相比，CVS有如下优点：1.支持Internet上的开发，而VSS只支持局域网上的开发；2.权限管理功能更强；3.支持多人同时Check Out一个文件；4.免费。

源代码版本管理软件虽然是为源代码管理而设计的，实际上可以用于任何的文本或数据资源的管理，特别适合于文本资源的管理。自然语言处理面对的是大量的文本，而CVS最适合于对文本并发编辑。用CVS就可以把代码资源、语言资源和文档资源都统一管理起来了。

5. 项目规划

平台的重要性总是通过平台之上的项目体现出来。我们已经规划了句法分析器和树库建设[5]的两个项目。

句法分析器的项目将是一个侧重开放源码的项目。我们已经有了一个小型的树库以及利用我们的句法分析器得到的对于该树库的召回率、准确率等功能指标。这可以作为一个优化的起点。

概率句法分析依赖于树库的规模，我们的句法分析器在进一步改进的过程中遇到的最大问题就是数据稀疏的问题。这个问题的根本解决方案也就是树库的建设了。所以，与句法分析器项目相伴而生的一个项目就是树库建设项目。这个项目将是一个资源建设项目。初步的设想是采用人机互助的办法。发挥网络的分布性和这个平台的协调功能，先把任务分解，然后再归并到平台上来。

除了以上这两个主要的项目，我们还将会把我们在以前开放机器翻译系统中积累的一些资源、文档也以一些小型的项目形式公开出来。另外，我们还打算设计一些项目，征集一些志愿者作为项目管理者进行开放。例如我们打算征集的一个项目就是要将近年来发表在“Computational Linguistics”上的所有文章摘要翻译出来。也希望大家提供更多、更好的建议。

6. 总结和讨论

开放式开发的好处已经在软件技术的各个领域得到了证明。自然语言处理开放平台的目标就是在本领域探索一条开放和协作的道路。我们首先把资源加以分类，对各自的属性加以分析。在此基础之上，我们设计了数据库和 CVS 结合的平台架构，并规划了一些启动项目。开放式开发的核心是人，网络只是提供了一种最佳的媒介。开放平台的长远发展需要众多项目的加入，需要好的思路的汇集。

自然语言开放平台的真正成功，取决于它能否吸引到足够的“人气”，能否不断地更新、不断的发展。真诚希望我国有志于自然语言处理的研究人员，特别是广大的学生，能主动关心这个平台，为这个平台的发展出一份力，共同促进我国自然语言处理研究水平上到一个新的台阶。

参考文献:

[1] Eric, S. Raymond. Cathedral and Bazzar. http://www.tuxedo.org/~esr/writings/cathedral-bazaar

[2] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313—330, 1993.

[3] 张华平，刘群. 基于N-最短路径方法的中文词语粗分模型：中文信息学报； 2002 年 16 卷 5 期(9 月) p. 77-p. 84.

[4] Kevin Zhang (Zhang Hua-Ping), Qun Liu (Liu Qun), Hao Zhang (Zhang Hao). Automatic Recognition of Chinese Unknown Words Based on Role Tagging: 19th International Conference on Computational Linguistics, First SigHan Workshop; 2002-9; 台北.

[5] 白硕，张浩. 角色反演算法. 软件学报，已录用，2002

[6] http://www.mandarintools.com

致谢 感谢计算所软件室自然语言处理组的李继锋、张华平、王树西、李素建、王长胜等，大家热烈的讨论促进了这项工作的开展，大家的宝贵意见都在文章中得到了体现。特别感谢张奕滔同学，作为平台的主要建设者，他做出了更为详细的设计并加以实现，付出了辛勤的劳动。感谢室主任程学旗老师的大力支持。

作者简介： 刘群（1966－），男，江西萍乡人，在职博士生，副研究员，主要研究领域为机器翻译，自然语言处理与中文信息处理；张浩(1978－)，，男，山西孝义人，硕士生，主要研究领域为自然语言处理；白硕(1956－)，男，辽宁辽阳人，研究员，博士生导师，主要研究领域为自然语言处理、网络安全

An Open source platform for Chinese NLP*

LIU Qun[1,2] HANG Hao[1] BAI Shuo[3]

1(Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

2(Institute of Computational Linguistics, Peking University, Beijing 100871China)

[3](National Administrative Center for Network and Information Security, Beijing 100031, China);

E-mail: liuqun@ict.ac.cn

Abstract: An Open source platform for Chinese language processing is presented in this paper. This is a platform that supports concurrent development of open source projects, including ordinary programming projects,

*Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030510&G1998030507-4.

corpus annotating projects and documents repository building projects. This platform is expected to become a resource center for Chinese NLP.
Key words: open source; resource platform; Chinese NLP