

自然语言处理开放资源平台^{*}

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)

[摘要]我国自然语言处理研究在一定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资源平台的设计方案。这个平台能够共享源代码、语料库、词典、学术论文等各种资源,并支持协作式的项目开发。随着参与者的增多和项目的发展,这个平台一定会为我国自然语言处理的研究提供有力的支持。

[关键词]开放源码;资源平台;自然语言处理

[中图分类号]H08**[文献标识码]**A**[文章编号]**1003-5397(2002)04-0050-07

An Open Resource Platform for Chinese NLP

Liu Qun, Zhang Hao, Bai Shuo

Abstract: The lack of public resources is a big problem in the research of Chinese Natural Language Processing (NLP). This paper suggests that the practice of open source code in software development can be used in the research of NLP. An Open Resource Platform for Chinese NLP (www.nlp.org.cn) is presented in this paper. This is a platform that provides various shareable resources for Chinese NLP, including source codes, lexicon, corpus and academic theses. The platform also supports concurrent development of NLP projects, including ordinary programming projects, corpus annotating projects and repository documents projects. This platform is expected to become a resource center for Chinese NLP.

Key words: open source code; resource platform; Chinese NLP

^{*} 本文工作受国家 973 项目支持,项目编号是:G1998030507-4 和 G1998030510。

[收稿日期] 2002-04-10

[作者简介] 刘群,男,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,男,硕士生,主要研究领域为自然语言处理;白硕,男,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。

零·引言

我国的自然语言处理研究在一定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(Raymond, 1997)。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。

本文提出采用类似 Linux 的开放源代码方式,建设一个自然语言处理的开放资源平台。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。

本文还将探讨建设一个面向中文的自然语言处理开放资源平台的若干问题,包括其目标、意义和组织形式、整体设计,以及平台之上的项目管理,并介绍该项工作目前的进展情况。

一 目标与意义

1.1 我国自然语言处理面临的问题

近年来,我国的自然语言处理研究取得了很大的进展。不过,一些深层次的问题也显得更加突出。自然语言处理研究的对象是人类语言。而这个对象体系庞大,从词法层次、句法层次、语义层次到语用层次,现象纷繁复杂,任何个人或研究小组都只能将研究精力集中在某个小范围内,而不可能面面俱到。然而,自然语言问题本身的复杂性又决定了自然语言处理的任何问题都是互相交织在一起的,任何一个问题都很难与其他问题完全割裂起来处理。所以,对于自然语言处理研究来说,一套公用的基础设施就变得非常必要。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。而这个问题在我国的自然语言处理研究中显得尤为突出。这主要体现在以下几个方面:

(1)缺少公用的语言资源。语言资源,包括词典、语料库、规则库等等,是自然语言研究不可或缺的工具。目前,英语的语言资源已相当丰富,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。其中比较著名的语言资源包括:北京大学计算语言学研究所开发的《现代汉语语法信息词典》;北京大学计算语言学研究所、人民日报社和富士通公司联合开发的《人民日报标注语料库》;董振东先生的《知网》;梅家驹先生的《同义词词林》。这些资源对中文信息处理的研究起到了极大的推动作用。不过,与英语相比较,我们可以得到的可共享资源还是要少得多。仅举一个简单的例子,汉语的人名识别问题非常重要,可是却没有一部公用的人名词典供大家研究之用。

(2)缺少公用的软件模块。一些公用的底层软件模块对于自然语言处理来说也是必不可少的。不用说词法分析、句法分析这样的复杂模块,一些更底层的简单模块,比如词典检索、汉字代码处理等等,都要耗去编程者很多的精力。这种状况极大地妨碍了我国自然语言处理研究的进展,一个明显的问题就是,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。

(3) 缺少公用的测试平台。重视评测,是目前自然语言处理研究的一个重要特点。公共的测试平台可以使大家的研究工作有一个互相比对的基准,避免在低水平上重复研究,而可以集中精力探索有突破性的新方法。目前国际上一些著名的评测,如 MUC、TREC、CoNLL 等,都极大地促进了相关领域的研究工作。汉语的自然语言处理研究中就很缺少这种公用的平台。例如汉语的词语切分,虽然研究已经非常多,但由于缺少公用的测试平台,大家的研究工作缺乏可以比较的基础。虽然在国家 863、973 项目的范围内都组织过一些评测活动,这些评测活动也都对相关的研究工作起到了很大的促进作用,但是由于这些评测的数据、程序、规范都没有做到完全公开,后续的研究工作无法沿用这些评测进行比较,这也使得这些评测的影响受到了一定的局限,还没有形成真正意义上的公共测试基准(benchmark)。

(4) 缺少公用的文献资料库。对于中国的研究者来说,虽然一般而言,阅读英语文献都不成问题。但是要比较全面的掌握相关研究的最新动态,要阅读大量的文献资料,还是要付出比西方学着多得多的精力。特别是对一些初次进入此领域的研究者来说,往往有点无从下手。建立一个比较完备的文献资料库,对于从事此领域研究的人来说,无疑是非常有益的。

(5) 缺少交流合作的机制。过分分散也是我国自然语言处理研究所面临的重要问题。当然这里有很多的客观原因,并不容易解决,例如缺乏经费的支持、单位之间的合作涉及知识产权问题等等。由国家相关管理部门(如自然科学基金委、863、973 专家组等)出面,统一牵头组织攻克一些大的研究课题,是一个好的解决办法。不过,对于这样一个涉及众多单位的大型研究课题来说,项目的组织管理工作是非常重要的。从另一个方面看,“开放源码”和 Linux 的成功,为我们指明了另一条可行的道路。

1.2 开放源代码的含义

“开放源码(open source)”的概念由公益组织“开放源码促进会(OSI)”(参考文献[6])所定义,对这类软件用户有使用、修改、复制的自由,因此这类软件的许可证(license)的宗旨往往与传统商业软件相反,旨在保护用户的使用权力。软件的作者乐意共享其劳动成果,也欢迎同行参与对该软件的批评与改进。

根据 OSI 的定义,一个“开放源码”的软件,该满足以下条件:(1) 自由重发布;(2) 提供源代码;(3) 允许再开发;(4) 原作者的代码完整性;(5) 没有对个人或群体的歧视;(6) 没有对应用领域的歧视;(7) 发布许可证;(8) 许可证不能针对某个产品;(9) 许可证不能限制其他软件。

“开放源码”的软件又称为“自由软件(free software)”。所谓的“自由”,也并不是完全没有限制。“开放源码”软件的传播通过一定的许可证来进行规范。有很多种不同的“开放源码”软件的许可证形式,其中最常见的是 GPL 许可证(参考文献[7])和 FreeBSD 许可证。

Linux 的成功已经证明,开放源代码是一种有效的软件开发方式。开放源代码不仅仅适合于小型软件,对于操作系统这样的大型复杂软件同样适用。实际上,开放源代码已经形成了一整套完整的软件开发模式,并有相应的工具软件(都是自由软件),可以支持互联网上众多的互不相识的人们共同开发一个完整的软件。

对于我们来说,开放源代码方式最具有吸引力的地方在于:通过这种方式可以组织众多分散的自然语言处理研究者、爱好者,利用他们的业余时间来做一些对于推动我国自然语言处理研究有益的事情。

1.3 自然语言处理开放资源平台的目标和意义

我们提出的自然语言处理开放平台和“开放源代码”还不完全是一回事。因为在我们的开放平台上,不仅仅有源代码,还有各种形式的资源,因此我们称之为“开放资源平台”。对于自然语言处理的资深研究者来说,我们希望为他们提供一个发布他们的某些(没有版权问题的)研究成果的场所;对于自然语言处理的入门研究者来说,我们希望给他们提供一个学习的环境,提供一个研究工作的起点;对于自然语言处理的业余爱好者来说,我们希望给他们提供一个与专业人士接触的机会和一个提高研究水平的途径;对于所有的自然语言处理研究者来说,我们都希望这是一个互相交流、共同提高的好地方。

二 开放资源的类型

建设一个开放资源平台,首先要将开放资源的类型加以明确。我们把自然语言处理的开放资源分为两种类型:一类称为静态资源,一类称为动态资源。

2.1 静态资源

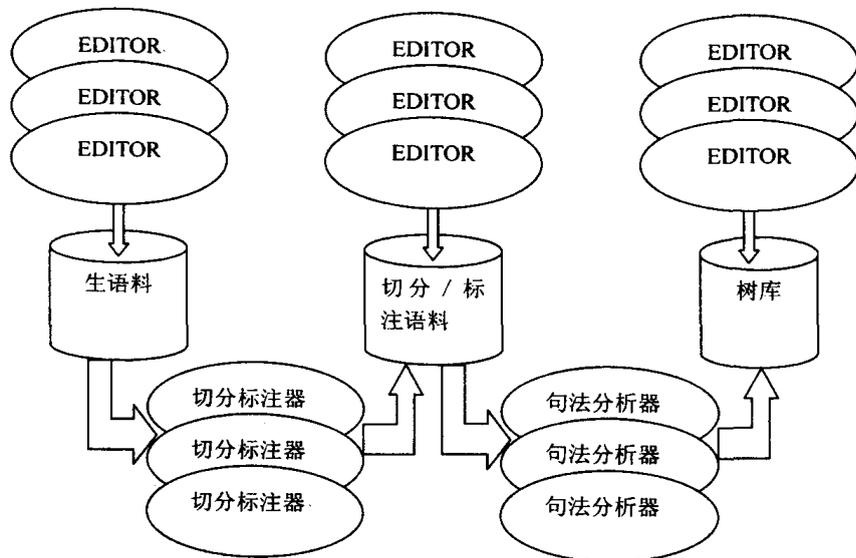
静态资源包括以下几类:(1) 源代码:目前各个领域都已有大量的开放源码计划。在中文信息处理领域方面,我们只在国外少数几个网站(Mandarintools, 参考文献[8])找到了很少的中文处理源代码,其中最复杂的是一个用 Perl 语言编写的汉语词法分析器,具有初步的词语切分和人名识别功能,正确率不高。其他方面几乎都还是空白。(2) 软件工具:各种以可执行程序形式(不含源代码)提供的应用程序和工具软件。(3) 词典:各种类型的词典是做计算语言学研究不可缺少的工具,如分词词典、人名地名词典、语义词典、拼音词典等等。现在真正可用的词典资源还是非常缺乏的。(4) 语料库:多种形式的语料库也是计算语言学研究的基础。如汉语切分标注语料库、语义标注语料库、双语语料库、树库等等。(5) 标准。(6) 论文:指正式发表的学术论文。(7) 技术报告:指正式发表的技术报告。(8) 技术资料(非正式发表):指非正式发布的技术资料,如各种课程讲义、学术报告、工程技术文档、技术规范等等。这些文档的重要性不亚于任何一种正式发表的论文和技术报告。推动自然语言处理作为一门学科的发展,迫切地需要各种形式知识的积累。实际上, Linux 下面的文档计划和源码计划的同步开展已经提供了成功的案例。(9) 对于文档类资源,开放平台应提供完善的管理和检索功能。(10) 网络链接:由于版权问题,我们不可能也没有必要把所有有用的资源都放在这个平台上。对于一些网上资源,可采用网络链接的方式提供给用户,并给出简短的文字说明。

2.2 动态资源

所谓动态资源,也称为项目,就是以前面所说的开放源代码形式进行组织的工程项目。项目的开发是一个动态的过程:人员上是动态变化的,时间上有起点和终点,并且按照项目开发的一般过程分为几个阶段。不过,我们这里的项目所开发的,不仅仅是一个软件,也完全可以其他的资源。例如语言资源(词典、语料库)、文档等等。下面我们通过两个例子来说明自然语言处理开放平台上的资源开放工作:(1) 《Computational Linguistics》论文摘要翻译项目。大量阅读论文是进行科学研究不可缺少的过程。不过由于英文水平和其他方面客观条件的限制,大多数国内的学者都很难像外国研究者那样掌握那么多的论文。即使对于一些英文程度较好的人来说,阅读英文文献的速度也大大低于阅读中文文献的速度。因此我们希望通过这个平台,组织一些项目,翻译一批经典学术论文的摘要。由于网上的人员层次较全面,可以找到各个领域的研究人员,因此这个工作由网上的虚拟研究小组来承担,甚至比任何一个具体的研究

小组更为合适。(2) 中文树库项目。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 (Marcus et al., 1993)。树库的开放需要耗费大量人力物力,其组织管理、规范制定、质量保证都是非常困难的。和开放源码的“Given enough eyeballs, all bugs are shallow”的思想 (Raymond, 1997) 相平行,我们认为,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。同时,吸收众多的研究者利用业余时间,每人贡献一点力量,也可以用一种低成本的方式开发出较大规模的树库。

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,如下图所示。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儿来调整的知识库。可以说,使规则系统完善的最好方法:将其开放,经受检验。



三 平台的组织形式

自然语言处理开放平台以网站形式呈现出来。

3.1 目录管理

为了访问者查找的方便,平台上所有的资源以领域分类目录的形式进行管理,同时提供站内搜索引擎,可以方便地进行检索。我们初步设计的领域分类目录形式如下:

- (1) 总论,包括:学术刊物,会议信息,好书推荐,网络资源;
- (2) 基础理论,包括:统计机器学习,汉语语言学;
- (3) 语言资源,包括:语料库,词典;
- (4) 关键技术,包括:汉字编码,词法分析,句法分析,语义分析;
- (5) 应用系统,包括:文本分类和聚类,信息检索和过滤,信息抽取,问答系统,拼音汉字转换系统,机器翻译。

3.2 用户管理

用户分为五类:网站管理员、领域负责人、项目负责人、普通注册用户、未注册用户。整个

网站设置一到多名网站管理员,负责整个网站的日常维护工作。对于领域分类目录中每一个领域,设置一到多名领域负责人,负责该领域资源的日常维护工作。领域负责人有整理资源、删除资源、建立子领域的权限。每个项目设置一个项目负责人,负责管理项目的开发工作。任何注册用户都可以申请设立一个项目并担任项目负责人,一旦项目被批准,就可以吸收其他注册用户加入项目并开始工作。普通注册用户可以浏览、下载、上载资源,参加项目。未注册用户只能浏览和下载资源,不能上载资源,不能参加项目。平台设立一个论坛,所有注册用户都可以在上面发表文章,进行交流。平台还提供一个邮件列表(mailing list)功能,用户可以按照自己的兴趣订阅邮件列表,通过邮件方式进行讨论。

3.3 项目管理 项目管理采用成熟的开放源代码的管理方式。利用版本管理软件实现开发人员之间的同步。

四 平台的实现方案

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环境与平台本身的操作系统环境无关,可以由项目任意指定。Web 服务器采用 Apache 服务器,动态页面通过 PHP + MySQL 的方式实现。平台上所有的静态资源都通过数据库 MySQL 进行管理。平台用户也通过 MySQL 进行管理,用户权限控制通过 Linux 本身的权限控制实现。MySQL 数据库中主要有以下几个数据表:用户数据表;学科数据表;资源数据表;项目数据表。

项目的管理较为复杂,主要通过代码版本管理软件 CVS 来实现。该软件用于具体的项目中所有文件的管理,可以实现文件的历史记录保存、版本比较、多人协同开放等等。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 就可以把项目中的代码资源、语言资源和文档资源都统一管理起来了。

站内搜索引擎、论坛和邮件列表都利用已有的自由软件,结合平台的具体需要定制而成。例如论坛的用户与平台本身的用户采用一套管理方式,用户加入论坛不必另外注册。这种定制的能力也是开发源代码给我们带来的方便。可以看到,整个平台都是在开发源代码软件的基础上实现的。

五 进 展

目前,自然语言处理开放资源平台已经开始试运行。我们已经为该平台注册了永久域名:www.nlp.org.cn。现阶段,平台的各项功能还不完善,还没有完全达到我们预定的目标,整个平台正在不断的完善过程之中。到目前为止,已有注册用户 40 人。平台上已有上载的静态资源 28 项,项目 2 个,分别是:词法分析器项目和概率句法分析器(白硕,2002)项目。这两个项目的初始源代码都由中科院计算所自然语言处理研究组提供。其中,词法分析器已经比较成熟,可以实现完整的词语切分、未定义词识别、词性标注功能,而且正确率很高(张华平等,2002; Zhang et al.,2002)。概率句法分析器也实现了一个功能相当完善的系统,只是由于训练语料

库较小,实验效果还不太令人满意。概率句法分析依赖于树库的规模,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。所以与句法分析器项目相伴,我们将会建立一个树库建设项目,初始树库是我们曾经开发的一个小规模树库,项目基本构思前面已经介绍过。除了这两个项目,我们还将会把我们以前开发机器翻译系统中积累的一些资源、文档也公开出来。另外,我们还打算设计一些项目,征集一些志愿者作为项目负责人进行开发。

六 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。自然语言处理资源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,我们提出了完整的平台设计方案并已基本实现。目前平台已经开通并试运行,已经上载了一批资源,并启动了两个项目。平台试运行时间不到两周,已经吸引了很多研究人员(包括海外研究人员)进行注册,并与其他一些专业站点实现了互相链接。

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. 1997. Cathedral and Bazaar. <http://www.tuxedo.org/~esr/writings/cathedral-bazaar>.
- [2] M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. 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[J].中文信息学报,2002年16卷5期,77-84.
- [4] Kevin Zhang (Zhang Hua-Ping), Qun Liu (Liu Qun), Hao Zhang (Zhang Hao). 2002. 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[J].软件学报,已录用.
- [6] OSI,开放源代码的定义,<http://opensource.org/docs/osd-sim-chinese.php>.
- [7] GNU,许可证,<http://www.gnu.org/licenses/licenses.cn.html>.
- [8] Mandarintools,<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。