LTP:语言技术平台

郎君，刘挺，张会鹏，李生

（哈尔滨工业大学信息检索研究室，哈尔滨 150001）

摘 要： 本文描述了一套面向 Web 基于 XML 的中文语言处理平台，命名为“语言技术平台 LTP”。LTP 包含 5 项主要内容：语言技术置标语言 LTML、基于 DOM Tree 的一套 DLL 模块、一套可视化工具、基于 LTML 的语料库资源、以及基于 Web Service 的网络应用。目前 LTP 集成了词法、词义、句法、语义、篇章分析等 10 项中文处理核心技术。该平台将为自然语言处理及信息检索领域的研究者提供一套系统化工具，帮助他们深入研究语言各个层面之间的关系并且利用这些基础技术去研究一些高级应用话题。

关键词： 自然语言处理；信息检索；XML；Web Service；技术平台

LTP: Language Technology Platform

Jun Lang, Ting Liu, Huipeng Zhang, Sheng Li

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001)

Abstract: This paper presents Language Technology Platform(LTP), the architecture of a Chinese processing platform for web application on XML. There are five main parts: Language Technology Markup Language(LTML), a suit of DLL modules based on DOM Tree, a suit of visualization tools, language corpora based on LTML and Web Service for LTP. It has integrated ten key Chinese processing modules on morphology, word sense, syntax, semantics and document analysis. A suit of systematism tools is supplied for beginners of natural language processing and information retrieval. They can study the relationship between levels and some advanced topics.

keywords: Natural Language Processing; Information Retrieval; XML; Web Service; Technology Platform

1 引言

任何技术的研究都要经历从综合到分析，再到综合的过程，人类语言技术也不例外。初期语言处理较简单，没有复杂的层次，比如机器翻译系统，只是查词典和简单调序。随着认识的深入，人们开始设计复杂系统，并针对各项单一技术进行深入的研究。经过几十年发展，语言技术在多项单一技术上取得了一定进展，但如果试图在不考虑其他层面的情况下去提升某一项技术的性能，就容易犯只见树木不见森林的错误。比如汉语分词发展到今天，语言模型等无法解决的一些歧义，往往和整句句法和语义相关。因此如果仍不考虑各层面技术间的照应而强求继续提高分词精度，理论上是无法突破的。语言技术领域，综合的时代正在到来。

本文组织结构如下：第二部分介绍语言技术平台的相关构想和涉及到的问题，第三部分介绍语言技术平台的构成和目前达到的水平，第四部分是平台的应用，第五部分是相关工作的介绍，最后是结论和展望。

基金资助：本文受到国家自然科学基金项目资助，项目号：60435020, 60575042, 60503072

作者简介：郎君（1981－），男，四川峨眉人，哈工大计算机系博士研究生，bill_lang@ir-lab.org

2 语言技术平台的构想

2.1 基本思想

语言理解是一个复杂的分层互动式系统，从以句子为处理单元的词法、词义、句法、句义、语用分析，到以篇章为处理单元的指代消解、自动文摘、文本分类，再到以篇章集合为处理单元的多文档文摘、文本检索等，构成了一个复杂的认知体系。我们认为：人类在理解语言时既不是简单的自底向上的分析，也不是简单的自顶向下的理解，而是以并行计算的方式，根据分析的需要，同时在各处理层面中穿梭、反馈、融合，找出最佳分析结果。

语言各个层面间的关系是错综复杂的，但一般来说，高层技术要建立在底层技术的基础上，同时又反过来指导底层技术。例如，“他从马上下来”有两个不同的分词结果。词汇化语言模型容易把应该拆开的两个词合并在一起，对此例就会错误地分为“他/从/马上/下来”。而利用高一层的词性信息，由于“从/马上”构成了“介词/副词”的邻接组合较少见，而“从/马/上”构成“介词/名词/方位词”的典型结构，因此容易得到正确的分词结果：“他/从/马/上/下来”。这是一个比较微观的案例，其实各层面间都存在着相互联系，只是联系方式和紧密程度有所不同。如：未登录词可能在文本中第一次出现时不易识别，但如果这个词在文本中多次出现，可以通过在整个篇章中计算汉字串频度的方式，把高频共现的汉字散串先找出作为未登录词的候选，然后再识别未登录词[1]。

在语言处理系统中，所谓高层和底层又是相对的，各项技术有一定的逻辑制约关系，又根据不同学者的不同认识和构思有相当的灵活性。词法分析在最底层，这一点大家一般都赞同，除非一些特殊场合直接用字作为处理单元。接下来，词义和句法孰先孰后就产生了争议。词义在处理单元上比句法小，但在理解深度上比句法深，因此句法可以利用词义消歧的结果构建基于名词语义类的句法分析，词义消歧也可以利用句法分析进行特征提取，找到和多义词真正有联系的上下文词，而不是简单开一个窗口[1]。在词义消歧和句法分析的基础上可以做浅层的语义分析[2]。句子级处理后是篇章级处理，技术间耦合变得更加松散，但仍然存在。比如有人基于文摘结果做文本分类，有人先体裁分类，再根据不同的体裁去生成不同的文摘。

2.2 错误级联的处理

分层处理，各层之间如果照应不好，最终效果不如不分层。因为分层后可能产生严重的错误累积，如系统分5层，每层的准确率为90%，简单串接在一起，最终准确率为60%。

解决错误级联通常有三种方法：一体化、逐级反馈、分层搜索。一体化把分布在不同层次上的多个结点捏合为一个结点，比如（词、词性）构成二元序对，在二元序对空间里搜索比分别在词网格和词性网格中搜索要复杂，但正是因为这种复杂才换来对各层信息的集成。逐级反馈是比较经济的方案，“眼前无路想回头”，有路就快速向前走，这和人的思维方式比较贴近，但困难之处在于如何找到反馈点。我们曾构造了一个汉语理解平台CUP(Chinese Understanding Platform)来尝试分层搜索[4]，每层都留下N个候选，构成庞大的搜索空间，试图在其中搜索出最佳路径，让各层面的知识在这条路经上通过竞争、平衡、融合。最终效果不佳，一个主要原因是句法分析准确率低，而候选空间太大，最佳的10个乃至于100个分析树都难以覆盖正确的结果。不准确的句法分析对分词、词性标注的指导意义没有体现出来。

在CUP上的尝试没有取得预期效果，但是我们认定在语言处理系统中分层是必需的，各层次之间的信息交流、融合也是必需的。为了深入分析各层次间关系，我们开发了语言技术平台LTP(Language Technology Platform)。

3 语言技术平台的主要内容

2006年4月28日，语言技术平台[5]发布，目前是中文处理的集成平台，囊括断句、分词、词性标注、命名实体识别、词义消歧、依存句法分析、浅层语义分析、指代消解、自动文摘和文本分类等10项中文处理技术。

LTP包含5项主要内容：LTML(Language Technology Markup Language)、基于DOM Tree的一套DLL模块、一套可视化工具、基于LTML的语料库资源、以及基于Web Service的网络应用。下面分五个部分来介绍。

3.1 LTML

综合的语言技术平台，需要一套清晰的数据内容表示方法，以及在此之上的各种相关处理和应用。我们基于XML 设计了一整套中文内部表示体系，从词处理到句子处理，再到篇章处理，直至篇章集合的处理，都能够用这套 XML 表示方法以一贯之地进行表示。这套表示方法我们称之为语言技术置标语言（LTML）。

3.1.1. LTML 的结构

为了清晰展示 LTML 的结构，我们举 2006 年 5 月 14 日新华网上的报道《"诗人"萨达姆静候绞刑架 聊天下大事论伊核危机》(http://news3.xinhuanet.com/world/2006-05/14/content_4543587.htm)为例。经过 LTP 处理生成的 XML 文件如图 1、2 所示。

```
<?xml version="1.0" encoding="gb2312" ?>
<?xml-stylesheet type="text/xsl" href="ltp_style.xsl" ?>
<ltml>
    <doc>
        <para id="0">
            <sent id="0" cont="英国《星期日泰晤士报》14日披露了伊拉克前总统萨达姆最近与一名女律师交谈的内容。">
            <word id="0" cont="英国" pos="ns" ne="S-Ns" wsd="Di02" parent="3" relate="ATT" />
            (部分省略)
            <word id="6" cont="披露" pos="v" ne="O" wsd="Hi14" parent="-1" relate="HED">
                <arg id="0" type="施事" beg="0" end="3" />
                <arg id="1" type="时间" beg="5" end="5" />
                <arg id="2" type="受事" beg="8" end="20" />
            </word>
```

图 1 LTML 结构（句子级信息）

Fig.1 The structure of LTML(sentence level)

```
    <class>军事</class>
    <sum>英国《星期日泰晤士报》14日披露了伊拉克前总统萨达姆最近与一名女律师交谈的内容。(部分省略)</sum>
    <coref>
        <cr id="0">
            <mention id="0" beg="196" end="196" />
            <mention id="1" beg="206" end="206" />
        </cr>
    </coref>
</ltml>
```

图 2 LTML 结构（篇章级信息）

Fig.2 The structure of LTML(document level)

图 1 是文件开头部分句子级的信息。其中 doc 表示篇章，para 表示段落，sent 表示句子，word 表示词语，arg 是浅层语义标注的谓词。各种 id 表示当前层面的节点编号。sent 中 cont 表示原句内容，word 中 cont 表示词条，pos 表示词性，ne 表示命名实体标签，wsd 表示词义消歧的义项代码(根据哈工大信息检索研究室同义词词林扩展版确定)，parent、relate 分别表示依存句法分析后当前节点的父节点 id 和发出弧线的关系类型。arg 所在的 word 是谓词， type 表示语义类型，beg 和 end 表示谓词约束范围。

图 2 显示文件的结尾部分，显示篇章级处理的结果。class 和 sum 分别表示目前文本的类别和文摘。coref 表示文本中代词和人名的指代关系，cr 表示文中的指代实体，下面的 mention 表示指代实体下面的各个指代对象，随后的 beg 和 end 表示在文本中绝对位置的开始和结尾。

3.1.2. LTML 的优缺点分析

我们曾设计过一套句子级的二维表结构，每一行是一个词，每一列是这个词的一个词法、语法或语义属性。当用这种结构描述篇章级信息，如指代消解时就遇到了困难，而如果试图去描写篇章集合的信息，比如多文档文摘，那就更困难了。以至长时间以来，句子级和篇章级的内部表示无法统一。现在 LTML 解决了这个问题。同

时 XML 是目前通用的表示手段，是语义网的基础。采用 XML 进行表示也有利于和未来的各种应用对接。

LTML 的另一优点是可以方便地将 XML 文件读到内存中进行操作，减少了以往文本表示和内存表示分离的情况。现在流行的 XML 操作库有 MSXML，Xerces-C++，TinyXML 等。

使用 XML 进行表示的一个缺点是，各种 tag 标记占用的存贮空间差不多是正文的 10 倍，对于需要高速处理的一个应用，比如互联网搜索而言，这是一个很大的麻烦。

3.2 基于 DOM Tree 的一套 DLL 模块

基于 TinyXML（http://www.grinninglizard.com/tinyxml/index.html），我们编写了一个 LTML 的操作函数库，包含各个自然语言处理模块的接口。LTP 目前包含的 10 个模块提供的都是 DLL。该框架将模块内部开发和外部调用完全分开，只要按照我们的 DLL 的接口方式，任何人都可以用新的模块来替代。

以往在开发应用系统的时候常常出现对底层模块重复调用的现象，比如在一套新闻搜索系统中，文本分类和自动文摘都调用了分词模块，造成了系统资源的浪费，降低了处理速度。在 LTP 中，我们把分词结果保存在 DOM Tree 中，其他模块都通过 DOM Tree 来调用分词结果，从而避免了重复调用问题。

3.3 处理结果可视化

一直以来，处理结果的可视化都困扰着集成化自然语言处理平台的建设。基于 XML 的 LTML 为结果可视化提供了天然基础。在网页技术中，XML 的可视化可以采用 Javascript，HTML，CSS，XSLT 和 VML 等技术实现。

在 LTP 中，输入一篇文本，经处理后，可以从不同角度、粒度去浏览处理的结果。通过对处理结果的对比分析，可以非常直观地看出各项结果间的关系，如图 3、4 所示。

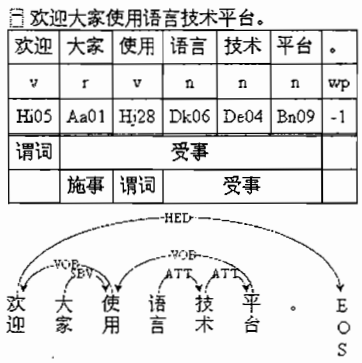

图 3 LTP 处理结果在 IE 上的部分显示

Fig.3 Some processed result of LTP displayed on IE

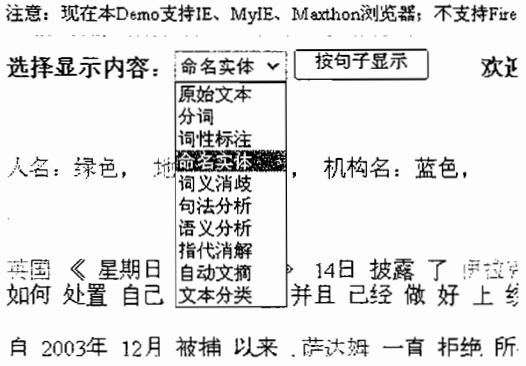

图 4 LTP 处理结果的多层面显示选择

Fig. 4 Multi-level choosing of processed result of

3.4 基于 LTML 的语料库资源

形成统一数据表示后，我们可以非常方便的将很多的自然语言资源 XML 化。到目前为止国内外已有 80 多家科研院所共享使用了我们的问答系统问题集、同义词词林扩展版等 6 种语料库。它们是 LTP 的重要组成部分，研究者可以用来开发和测试相关模块。但是目前它们没有统一的格式，在使用时需要根据具体的数据来进行使用。我们准备将它们按照 LTML 重新进行规范化。这对于我们的科研及对外共享都会很方便。

3.5 以 Web Service 的方式提供技术服务

现在的互联网是一个提供“内容”和“服务”的时代，大量出现的 Web Service API 使我们可以方便地将自己的应用和别人的成果链接起来。比如 Google 有 Search API， Maps API 等；Baidu 提供了 Search API；Yahoo 提供了天气资讯的 API。

Web Service 的一个关键问题就是结果的 XML 化，目前 LTP 处理生成的 XML 文件恰好满足这种需求。我们可以实现各种自然语言处理模块的 Web Service。对于一些自然语言处理之外的系统，当需要相关的自然语言处理功能时可以非常方便的通过网络调用的方式来完成。LTP 的 Web Service 将会是一种很好的应用。封装好各种调用接口之后，我们可以完全关注在各种 NLP 技术的不断完善和提高上，使用 LTP Web Service API 接口的系统

也可以更加关注于具体应用。Web Service 还可以不断的将各种实际应用中模块的输入和输出显现在研发人员面前，这对于具体模块的开发会形成极大的促进。

4 应用

总的来说，LTP 试图解决语言处理的多层次（句、篇，文本集合）机内表示问题，语言处理结果的可视化问题，以及各处理模块避免重复调用的问题，能够帮助 NLP 研究者从更高的起点直接进入高层技术的研究。

LTP 平台可以用于 NLP 技术的教学。在 LTP 网站上，输入一段文本，可以非常方便的看到各个层面的处理结果。对于 NLP 领域的初学者，可以方便的学习和理解各种自然语言处理技术，在我们共享的各种模块的基础上也可以很好的搭建自己的实验平台。

5 相关工作

由于英文上研究自然语言处理比中文要早，国外的自然语言处理的平台有很多。其中比较著名的系统有 GATE、NLTK 和 NLPWin 系统。

GATE(General Architecture for Text Engineering)，是英国谢菲尔德大学自然语言处理组开发了 11 年的一个自然语言处理平台[6]，包含统一的开源体系结构和图形化开发环境。这个平台基于 XML 进行数据表示和处理，集成了大量自然语言处理资源，被用来进行自然语言处理的相关教学和研究。NLTK(Natural Language Toolkite)[7]是自然语言工具包，是一套用于自然语言处理的符号和统计处理的 Python 程序库。NLTK 包含图形化的演示和样本数据。它包含一整套扩展的文档，包含支持这套工具集的自然语言处理的相关概念的解释。NLPWin[8]是微软研究院在 90 年代开发了一个通用型的 7 国语言处理平台。其中文部分采用"切词-句法分析一体化"方法。

信息检索研究室以往开发过中文理解平台 CUP（Chinese Understanding Platform）[5]。现在的 LTP 也部分借鉴了当时的经验和教训。

6 结论和展望

目前 LTP 已经包含五大部分。我们的目标是把 LTP 建设成 GATE 那样的中文方面的语言技术平台，包含相关的技术研究，语料资源，各种应用，以及促进中文自然语言处理的教学。

下一阶段，我们会从以下方面继续完善 LTP：进一步规范化 LTML 的表示形式；集成复合名词短语识别、文本聚类和多文档自动文摘等技术模块；采用 Web Service 的方式对外提供核心技术服务；不断地提高每个单项技术的精度；逐步尝试采用新的复杂架构对各项技术进行整合，通过反馈、搜索等各种机制提高系统的整体性能。

参考文献：

[1] 刘挺, 吴岩, 王开铸. 串频统计和词形匹配相结合的汉语自动分词系统[J]. 中文信息学报,1998,(1).

[2] 卢志茂, 刘挺, 张刚, 李生. 基于依存分析改进贝叶斯模型的词义消歧[J].高技术通讯,2003(5)

[3] 车万翔, 刘挺, 李生. 浅层语义分析[A]. 全国第八届计算语言学联合学术会议(JSCL-2005)论文集.2005 年 8 月,南京

[4] Wanxiang Che, Ting Liu,and Sheng Li. A New Chinese Natural Language Understanding Architecture Based on Multilayer Search Mechanism[A]. Third SIGHAN Workshop on Chinese Language Processing (SIGHAN2004), pages 134-140, Aug. 2004

[5] http://ltp.ir-lab.org

[6] Cunningham, H., D. Maynard, K. Bontcheva and V. Tablan. Gate: A Framework and Graphical Development Environment for Robust Nlp Tools and Applications[A]. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics 2002.

[7] Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit[A]. Proceedings of the ACL demonstration session, Barcelona, Association for Computational Linguistics, pp 214-217. July 2004.

[8] Knight, Kevin and Vasileios Hatzivassiloglou. Two-Level, Many-Paths Generation[A]. Proc. Conf. Assoc. for Computational Linguistics (1995): 252-60.