基于 XML 的语言技术平台

李正华，车万翔，刘挺
哈尔滨工业大学计算机学院 信息检索研究中心 哈尔滨 150001
Email: {lzh, car, tliu}@ir.hit.edu.cn

摘 要：本文提出了一个基于 XML 数据表示的中文自然语言处理框架：语言技术平台（Language Technology Platform, LTP）。LTP 包括六个组成部分：基于 LTML (Language Technology Markup Language) 的数据表示方法、基于 LTML 的语料库资源、面向中文的语言处理模块、基于动态链接库（DLL）的应用程序接口、可视化工具、以及网络服务 Web Service。LTP 采用了分层的结构，涵盖了词法分析、句法分析以及语义分析等六项语言处理任务。LTP 免费向学术界共享，很多研究单位已经在 LTP 基础上做出了科研成果。
关键词： 语言技术平台；语言处理模块；可视化；语料资源

XML-based Language Technology Platform

Zhenghua Li, Wanxiang Che, Ting Liu
Research Center for Information Retrieval of Computer Science & Technology School, Harbin Institute of Technology, Harbin 150001
Email: {lzh, car, tliu}@ir.hit.edu.cn

Abstract: This paper presents an XML-based framework of Chinese processing platform, named Language Technology Platform (LTP). LTP consists of six components: XML-based data representation named Language Technology Markup Language (LTML), a suit of Chinese processing modules, a suit of API which is implemented as dynamic link library (DLL), visualization tools, a set of corpora resources, and web service. A layered architecture is used to integrate six key Chinese processing modules on morphology, syntax, semantic, and document analysis. LTP is shared by the public for academic purpose. It has benefited many researchers, who have done excellent work based on LTP.
Keywords: Language Technology Platform; Language Processing Module; Visualization; Corpora Resource

1 引言

随着自然语言处理的研究逐渐深入，各种形式的语义分析成为当前的研究热点。然而，目前句法分析、语义分析等深层的自然语言处理技术没有真正支持实际应用。如信息检索对自然语言处理技术的使用还只是停留在基于分词结果的阶段。究其原因，自然语言处理系统效率不高、准确率偏低固然是障碍其实际应用的主要原因。另外一个不可忽视的因素在于自然语言处理领域的入行门槛较高，加之缺少共享的数据和程序资源，使得从事应用领域的研究人员很难快速构建一个可用的自然语言处理系统，这也导致其应用自然语言处理结果的兴趣和信心大打折扣。

另一方面，就自然语言处理领域的研究人员而言，由于缺乏数据和相应的基础模块，重复制造轮子的事情也屡有发生，这一定程度上造成了人力和资源的浪费。同时，由于缺乏分析结果的可视化工具，语言分析一直被看成是黑箱，错误分析成为一个难点，这不利于对数据的观察，也就很难产生语言上的直觉。

为了解决以上提到的问题，我们建设了语言技术平台 LTP（Language Technology Platform）。LTP 使用 XML 作为底层数据表示，提供了丰富、高效的中文处理模块，基于

基金资助：国家自然科学基金项目（60803093；60975055）；国家 863 项目（2008AA01Z144）
作者简介：李正华（1983-），男，博士研究生，依存句法分析；车万翔（1980-），男，博士，讲师，自然语言处理；刘挺（1972-），男，教授，博士生导师，信息检索和自然语言处理。

动态链接库（Dynamic Link Library, DLL）的应用程序接口，可视化工具，语料库资源，并能够以网络服务（Web Service）的形式进行使用。为了促进自然语言处理研究的发展，我们免费将 LTP 共享给研究界。迄今为止，国内外很多研究机构基于 LTP 发表了学术成果。

本文组织结构如下：第二部分介绍相关工作，第三部分介绍语言技术平台的建设情况，第四部分介绍语言技术平台的共享情况，第五部分是结论及下一步工作规划。

2 相关研究工作

自然语言处理平台的开发一直是很多关注应用的研究人员的目标。由于自然语言处理研究在英文上比中文的要早，国外的自然语言处理平台有很多。其中比较著名的系统有 GATE、UIMA 和 NLTK。

GATE（General Architecture for Text Engineering）[1]，是英国谢菲尔德大学自然语言处理组开发的自然语言处理平台，包含一个统一的基于 Java 的开源的体系结构和图形化的开发环境[1]。GATE 采用了基于组件的体系结构，语言处理、语料、及可视化资源都被表示为组件，从而可以促进资源的重用。GATE 提供了大量可重用的组件，被用来进行自然语言处理的相关教学和研究。另外，GATE 提供了一组集成的图形化工具，帮助使用者建立、修改和调试各种资源。

UIMA（Unstructured Information Management Architecture）[2]是一个用于开发、部署非结构化信息管理应用的软件架构[2]。它通过对文本、视频、音频、图片等非结构化信息的内容进行分析和组织，从而获取相关知识，产生结构化的、易于获取的数据，交付给终端用户使用。分析技术包括：基于统计的、基于规则的自然语言处理技术，信息检索、机器学习、本体知识、自动推理等。UIMA 和 GATE 类似，都采用了基于组件的设计模式，将语言处理核心算法和其他系统服务如数据存储、组件间通信、结果可视化分离。UIMA 强调对已有技术的利用、可扩展性、中间件和平台无关性。

NLTK（Natural Language Toolkit，自然语言处理工具包）[3]是一套用于自然语言处理的 Python 程序库[3]。NLTK 包含图形化的演示和样本数据。它还包含一整套扩展文档，支持这套工具集在自然语言处理中相关概念的解释。NLTK 被广泛应用于自然语言处理的教学和研究中。

以上平台的一个共性问题是它们都强调系统的体系结构，但缺乏精准的语言分析模块，尤其是中文分析模块。因此，有必要开发一套针对中文的自然语言处理平台。

3 语言技术平台

2006 年 4 月，哈工大信息检索研究中心推出了语言技术平台（Language Technology Platform, LTP）[4]。LTP 是一个中文处理的集成平台，囊括了词法分析（包括分词、词性标注和命名实体识别）、句法分析（依存句法分析）、语义分析（词义消歧和语义角色标注）等 6 项语言处理关键技术。其系统框架如图 1。

LTP 包含 6 项主要内容：基于 LTML(Language Technology Markup Language)的底层数据表示、基于 LTML 的语料库资源、面向中文的语言处理模块、基于动态链接库（DLL）的应用程序接口、可视化工具、以及网络服务（Web Service）。由于网络服务还处于建设阶段，下面分五个部分来介绍。

---

[1] http://gate.ac.uk/
[2] http://www.research.ibm.com/UIMA/
[3] http://www.nltk.org/
[4] http://ir.hit.edu.cn/demo/ltp

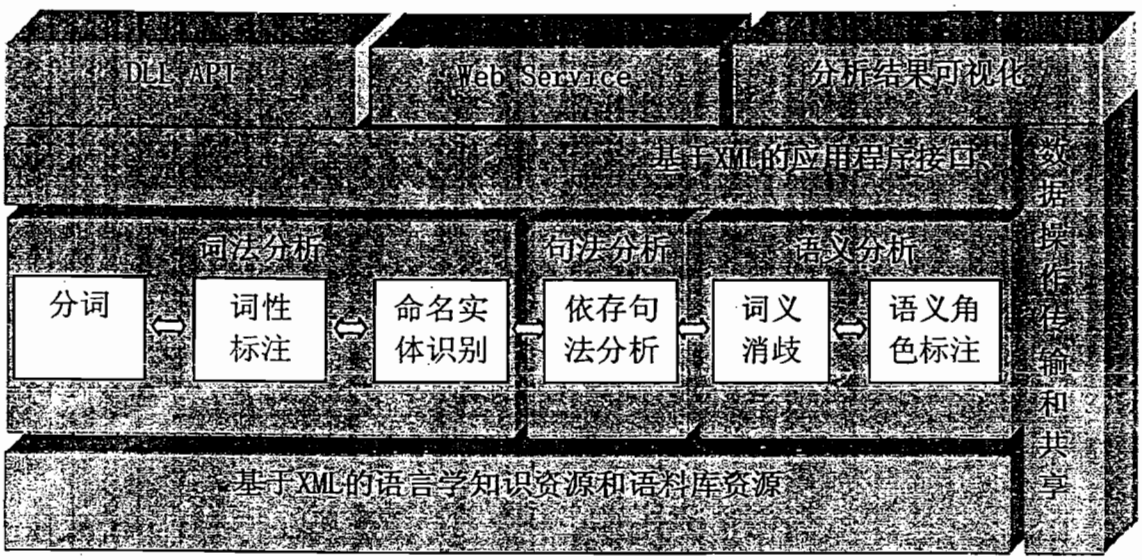

图 1 语言技术平台系统框架

3.1 数据表示

综合的语言技术平台，需要一套清晰的数据表示方法，以及基于这套表示方法的各种相关处理和应用。XML 作为一种清晰的数据表示方式，已经被大家所接受，并且逐渐成为一种标准的数据表示方式。基于 XML 我们设计了一整套中文内部表示体系，从词处理到句子处理，到篇章处理，直至篇章集合的处理，都能够用这套 XML 表示方法进行表示。这套表示方法我们称之为语言技术置标语言 LTML（Language Technology Markup Language）。

```
<?xml version="1.0" encoding="gb2312" ?>
<ltml>
    <doc>
        <para id="0">
          <sent id="2" cont="国内专家学者 40 余人参加研讨会。">
            <word id="0" cont="国内" pos="nl" ne="O" wsd="Cb05" parent="1" relate="ATT" />
            <word id="1" cont="专家" pos="n" ne="O" wsd="Al02" parent="2" relate="ATT" />
            <word id="2" cont="学者" pos="n" ne="O" wsd="Al01" parent="6" relate="SBV" />
            <word id="3" cont="40" pos="m" ne="B-Nm" wsd="-1" parent="5" relate="QUN" />
            <word id="4" cont="余" pos="m" ne="I-Nm" wsd="Dn05" parent="3" relate="RAD" />
            <word id="5" cont="人" pos="n" ne="E-Nm" wsd="Aa01" parent="6" relate="SBV" />
            <word id="6" cont="参加" pos="v" ne="O" wsd="Hj20" " parent="-1" relate="HED">
                <arg id="0" type="Arg0" beg="0" end="2" />
                <arg id="1" type="Arg0" beg="3" end="5" />
                <arg id="2" type="Arg1" beg="7" end="7" />
            </word>
            …
    </doc>
```

图 2 LTML 示例

我们以图 2 为例，详细介绍 LTML 的结构（如表 1）。其中<doc>表示篇章，<para>表示段落，<sent>表示句子，<word>表示词语，<arg>表示浅层语义角色标注的结果。各种节点下的 id 表示当前层面上的节点编号。sent 节点中的 cont 表示原句内容，word 节点中的 cont 表示词语内容，pos 表示词性标注结果，ne 表示命名实体标注的结果，wsd 表示全文词义消歧的词语义项代码，parent 表示依存句法结构中当前节点的父节点 id，relate 表示依存句法结构中在当前节点和父节点的关系类型。arg 所在的 word 节点表示当前 word 是一个谓词，arg 节点下的 type 表示语义类型，beg 和 end 分别表示论元的位置范围。

表 1 LTML 标记和属性的含义

Tag	说明	属性	说明
<doc>	篇章		
<para>	段落	id	篇章中段落的编号
<sent>	句子	id	段落中句子的编号
<word>	词语	id	句子中词语的编号
		cont	词语的内容
		pos	词性
		ne	命名实体
		wsd	《同义词词林(扩展板)》的词义代码
		parent	依存句法结构中父节点的编号
		relate	依存关系类型
<arg>	谓词的论元	id	论元的编号
		type	论元的内容
		beg	论元的开始位置
		end	论元的结束位置

各种编程语言，都提供了丰富的 XML 操作库。LTML 作为语言技术平台的底层数据表示，对各个模块之间进行信息传递、信息融合以及最终结果的可视化都提供了诸多便利。

3.2 语料资源

我们对外共享了 2 种与 LTP 相关的语料库资源。详细情况如表 2 所示，这些语料库是 LTP 的重要组成部分。

表 2 LTP 语料资源

语料库名称	规模	说明
同义词词林扩展版	77,343 条词语(原《同义词词林》[4]3 万余条)	秉承《同义词词林》的编撰风格，同时采用五级编码体系
中文依存树库[5]	不带句法关系 5 万句 带句法关系 1 万句	LTML 化，分词、词性、句法部分人工标注，可以图形化查看

3.3 语言处理模块

LTP 提供了 6 个中文处理模块。这些模块均采用了当前比较成熟的方法实现，兼顾了效率和性能。分别如下：

- 分词(Word Segmentation)

  该模块具有分词、时间数词识别、未登录词识别等功能。系统采用 CRF 模型(Conditional

Random Field)[6]，利用 1998 年上半年人民日报语料中的 1-5 月份训练，6 月份语料测试，精确率、召回率、F 值分别达到 97.2%、97.7%、97.4%。效率约为 185KB/s。

- 词性标注 (POS Tagging)

采用 SVMs(Support Vector Machines)模型，利用 1998 年上半年人民日报语料中的 2-6 月份训练，1 月份语料测试，基于正确的分词结果，整体准确率为 97.80%，未登录词(未出现在训练数据中的词)准确率 85.48%。效率约为 56.3KB/s[7]。

- 命名实体识别 (Named Entity Recognition)，NER

该模块可识别人名、地名、机构名、专有名词、时间、日期、数量短语等七类实体。该模块采用统计和规则相结合的方法实现。先利用最大熵(ME)的方法对文本初始标注[8]，然后再利用规则的方法对错标或漏标的结果进行修正。利用 1998 年 1 月份人民日报语料训练，共 37,422 句，1998 年 6 月份人民日报语料前 10,000 句测试，总的 F 值为 92.3%。效率约为 14.4KB/s。

- 词义消歧 (Word Sense Disambiguation)，WSD

该模块采用《同义词词林(扩展版)》的词义标注体系，能够对给定文档中的全部词汇进行词义标注。这里使用前三层编号，因为前三层标号已经能够区分大部分多义词了。模块采用基于 SVMs(Support Vector Machines)模型的有指导词义消歧方法[9]。利用哈工大信息检索研究中心标注的全文词义消歧语料库前 8,000 句作为训练集，8,000-9,000 句作为开发集，最后 1,000 句作为测试集。基于正确分词，多义词准确率为 91.29%，全部词为 94.34%。效率约为 7.2KB/s。

- 依存句法分析 (Dependency Parser)，Parser

依存句法分析系统用于对中文进行句法分析，将句子由一个线性序列转化为一棵结构化的依存分析树，通过依存弧反映句子中词汇之间的依存关系。该模块采用基于最大生成树算法，使用了高阶的特征。该模块参加了 CoNLL2009 多语种（包括中文、英文在内的 7 种语言）依存句法分析和语义角色标注评测，在 21 家参赛单位中获得句法分析第 3 名[10]。LTP 中的句法分析模块使用哈工大信息检索研究中心自行标注的依存树库前 8,000 句作为训练集，8,000-9,000 句作为开发集，最后 1,000 句作为测试集。基于正确分词、自动词性标注，该模块的依存关系准确率（LAS）为 73.91%，依存弧准确率（UAS）为 78.23%。效率约为 0.2KB/s。

- 语义角色标注 (Semantic Role Labeling)，SRL

语义角色标注任务是识别句子中的谓词以及该谓词的语义角色。该模块采用基于最大熵的统计方法实现，同样参加了 CoNLL2009 评测，最终获得了第 1 名[10]。其中基于自动依存句法分析，中文语义角色标注的性能为 F=77.2%。效率约为 1.3KB/s。

3.4 应用程序接口

基于 TinyXML，编写了一个 LTML 的操作函数库，包含基本的 XML 操作功能和相关的各个自然语言处理模块的接口，从而将 LTML 结构和各个语言处理模块连接起来。LTP 提供了大量的接口，以方便用户的使用。接口大致可以分为三类：IO 操作、语言处理模块调用、处理结果抽取。

- IO 操作

主要负责将文本、字符串或者 XML 文件读入，在内存中转化为 DOM 结构；或者将内存中的 DOM 写成 XML 文件。

- 语言处理模块调用

LTP 中提供的 6 种语言处理模块对应 6 个接口。

- 处理结果抽取

用户通过这些接口获得各个语言处理模块的结果。

目前 LTP 提供两种语言的接口，分别为 C++和 Python。程序包通过配置文件的方式确定每个模块的参数配置、数据文件的路径等。目前 LTP 程序包只支持在 Windows 操作系统下运行。

3.5 处理结果可视化

清晰的处理结果可视化可以帮助研究人员方便的进行错误分析等各项工作。我们在 LTML 的基础上，开发了一套基于 Flex 的可视化工具，用户可以在不同操作系统和浏览器上进行显示。一篇文本经过 LTP 处理后，可以从不同角度、粒度去观察处理的结果。

图 3 显示的是所有句子级模块的处理结果。其中第一行为分词信息；第二行为词性信息；第三行为词义信息；第四行为命名实体信息；第五行之后为语义角色标注结果，每一个谓词占一行。最上面的弧表示依存分析结果。

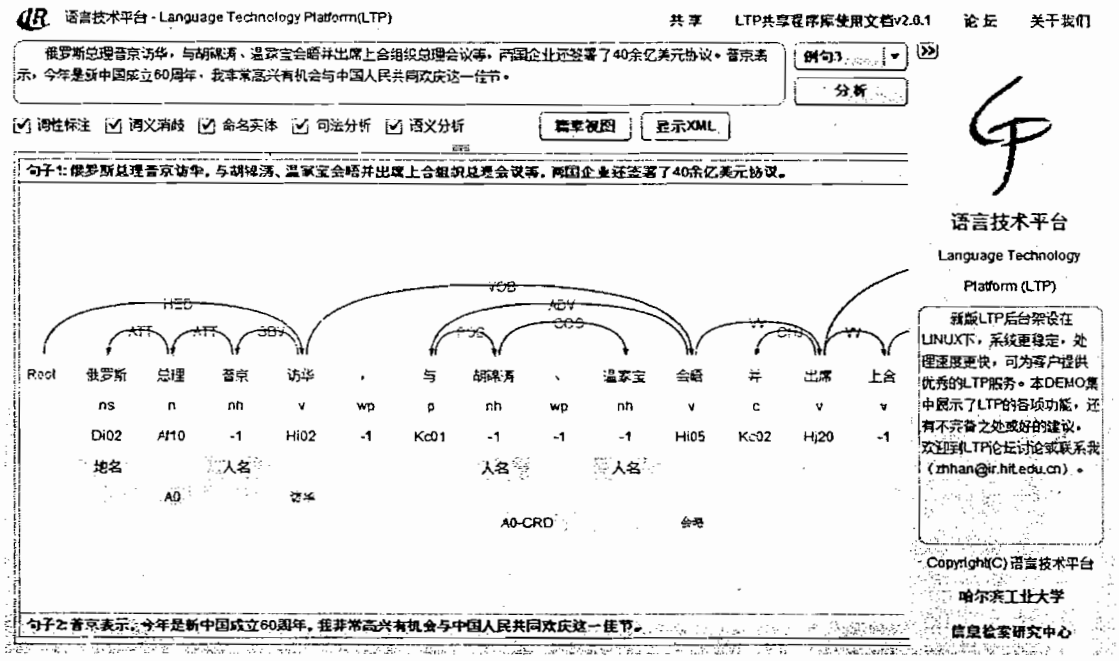

图 3 LTP 句子级模块可视化结果

4 语言技术平台升级及共享情况

我们一直在不断的完善语言技术平台。各个语言处理模块的性能不断的提高，整个系统架构在不断的优化，可视化程序不断的完善。截止 2009 年 9 月，LTP 已经升级至 2.0 版本。

为了进一步促进中文信息处理的研究，尽可能为大家提供一个方便直接进入高层研究的语言处理平台，我们于 2006 年 9 月开始对学术界免费共享整套 LTP[1]。

截止 2010 年 6 月，共享单位达到 340 多家，包括国内外众多大学及科研机构。很多单

[1] http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

位已经在 LTP 的基础上进行研究并且发表论文，据不完全统计，目前基于 LTP 发表的论文超过 60 篇。

5 结论

我们提出了一个基于 XML 数据表示的自然语言处理框架：语言技术平台 LTP。在这个框架下，自然语言处理的很多任务被结合起来，包括词法分析、句法分析、语义角色标注等。我们定义了一套基于 XML 的语言技术置标语言（LTML）作为底层数据表示；我们提供了一套中文处理模块，基本涵盖了各种句子级任务；为了方便研究人员观察处理结果，我们开发了一套基于 Flex 的可视化工具；另外还提供了 2 种语料资源。我们免费向学术界共享 LTP，很多研究单位已经在 LTP 基础上做出了科研成果。

语言各个层面之间的关系是错综复杂的。但一般来说，高层的技术要建立在底层技术的基础上，同时又可以指导底层技术。目前为止，LTP 只是一个分层的语言处理过程，各层之间没有任何反馈或者信息传递。下一步我们将在 LTP 上尝试各种互动反馈机制，如一体化、重排序等策略，从而提高整个语言处理系统的性能。

参 考 文 献

[1] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an Architecture for Development of Robust NLT Applications [C]. Proceedings of ACL 2002, 168-175.

[2] David Ferrucci, Adam Lally. 2004. Building an Example Application with the Unstructured Information Management Architecture. IBM Systems Journal, 2004, 43(3):455-475

[3] Steven Bird and Edward Loper. NLTK: The Natural Language Toolkit [C]. Proceedings of the ACL demonstration session 2004, 214-217.

[4] 梅家驹、竺一鸣, 高蕴琦, 高鸿翔编. 同义词词林 [M]. 上海. 上海辞书出版社. 1983.

[5] Ting Liu, Jinshan Ma and Sheng Li. 2006. Building a Dependency Treebank for Improving Chinese Parser [J]. Journal of Chinese Information Processing. 2006, 16(4): 207-224.

[6] John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. Proceedings of ICML 2001, 282–289.

[7] 王丽杰, 车万翔, 刘挺. 2009. 基于 SVMTool 的中文词性标注 [J]. 中文信息学报, 2009, 23(4):16-21.

[8] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language [J]. CL 1996, 22(1):39-71.

[9] Yuhang Guo, Wanxiang Che, Yuxuan Hu, Wei Zhang and Ting Liu. HIT-IR-WSD: A WSD System for English Lexical Sample Task [C]. SemEval 2007.

[10] Wanxiang Che, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin and Ting Liu. 2009. Multilingual Dependency-based Syntactic and Semantic Parsing [J]. Proceedings of CoNLL 2009, 49-54.