借助汉-越双语词对齐语料构建越南语依存树库*

李发杰[1,2]，余正涛[1,2]，郭剑毅[1,2**]，李英[1,2]，周兰江[1,2]，毛存礼[1,2]

(1.昆明理工大学 信息工程与自动化学院，云南 昆明 650500；

2.昆明理工大学 智能信息处理重点实验室，云南 昆明 650500)

摘 要： 由于对越南语的研究工作相对比较少，因此还没有建立规模相对较大的依存树库。相对于已经拥有了形态丰富、语料成熟的汉语，越南语的依存句法分析要困难的多，所以本文提出了一种借助汉—越双语词对齐语料构建越南语依存树库的方法。首先对汉语—越南语句子对进行词对齐处理，然后对汉语句子进行依存句法分析。最后结合越南语本身的语言特点和有关的语法规则将汉语的依存关系通过汉—越双语词对齐关系映射到越南语句子中，从而生成越南语的依存树库。实验表明，该方法简化了人工收集和标注越南语依存树库的过程，节省了人力和构建树库的时间。实验结果表明，该方法相比采用机器学习的方法准确率明显提高。

关键词： 越南语依存树库；汉语依存句法分析；汉—越语言对齐关系

中图分类号：TP391 **文献标识码：**A

Built Vietnamese Dependency Treebank By means of Chinese-Vietnamese Bilingual Corpus of Word Alignment

Li Fajie[1,2],Yu Zhengtao[1,2], Guo Jianyi[1,2],Li Ying[1,2], Zhou Lanjiang[1,2], Mao Cunli[1,2]

(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500，China

2. The Key Laboratory of Intelligent Information Processing,Kunming University of Science and Technology, Kunming, Yunnan 650500，China)

Abstract Few studies on Vietnamese, therefore has not built relatively large dependency Treebank. Compared to the rich and mature Chinese corpus, Vietnamese Syntactic Analysis is much more difficult. This paper presents an approach of Han-Vietnamese bilingual corpus of word alignment built Vietnamese Dependency Treebank method. Firstly, the aligned word processing was made by Chinese-Vietnamese sentence pairs；Secondly, the dependency parsing was done with Chinese sentences. Finally, Vietnamese Dependency Parsing Treebank was generated by Chinese-Vietnamese Languages align relationship and Chinese Dependency Tree. Experimental results show that this approach can simplify the process of manual collection and annotation of Vietnamese Treebank, also can save manpower and time building the Treebank. Experimental results show that the accuracy of this method compared to machine learning methods has improved significantly.

Keywords: Vietnamese Dependency Treebank;Chinese Dependency Parsing;Word Alignment

收稿日期：2015-5-31 **定稿日期：**2015-8-10

***基金项目：**国家自然科学基金(61262041, 61472168)；云南省自然科学基金重点项目(2013FA030)

**通讯作者：gjade86@hotmail.com

1 引言

越南与云南山水相连，两国人民之间的交往历史悠久，语言沟通在双方人民友好往来与相处、相互学习方面起到了十分重要的作用。因此，针对汉越双语的研究工作具有重要的现实意义。在越南语和汉语的互译过程中，越南语的句法分析是十分重要的基础工作。完全句法分析要求通过一系列分析过程，最终得到句子的完整的句法树；而浅层句法分析不要求得到完全的句法分析树，只要求识别其中的某些结构相对简单的成分，即它将句法分析分解为两个子任务：语块的识别和分析；语块之间的依附关系分析。由于采用完全句法分析难度相对比较大，因此浅层句法分析成为当前句法分析主流[1]。依存句法分析是机器分析语言句法特征非常有效的方法之一，本文对越南语采用依存树的方法进行句法分析。越南语依存标注体系和越南语依存树库的构建，已经成为整个越南语依存分析的核心工作，对该问题加以有效合理的解决，对越南语的句法分析、机器翻译、信息获取等上层应用可以提供有力支撑。依存句法分析的研究工作以及依存树库的建设工作，在国内外都已经开始展开。比较著名的依存树库有：捷克语的布拉格树库[2]，英语的PARC树库[3]，以及俄语、意大利语等语言的树库[4][5]。在中文方面也建立了一些比较有影响力的依存树库，如HIT-CIR-CDT，哈工大社会计算与信息检索研究中心汉语树库[6]，120万词，6万句子。在越南语的依存树库建设方面，P.T.Nguyen等人开展了依存树库的构建工作[7]，但其规模较小，共1万个句子左右，不能满足汉越双语机器翻译的需求。

从以上分析可以看出，大语种树库的建设工作已取得了一些成果，但由于对越南语而言，其研究工作相对比较少，还缺乏一定规模的依存树库。越南语与中文一样，已经标记好的依存句法树库资源是统计依存句法结构分析必备资源，如何实现构建越南语的依存树库也成为本文工作主要解决的问题。

本文针对越南语言特点，提出了借助汉-越双语词对齐语料构建越南语依存树库的方法，实验结果表明：本文提出的方法相比采用机器学习的方法在依存弧准确率（Unlabeled Attachment Score,UAS）、标识准确率（Labeled Attachment Score，LAS）和根节点正确率（Root Accuracy，RA）都有一定的提高。

2 汉越两种语言之间的差异

经过对越南语和汉语的对比研究发现，两种语言在语法结构上存在一些差异：（1）越南语定语位置和汉语不同，越南语定语一般在中心词后边，例如，汉语“她是美丽的女孩。”，越南语“Cô là một (她是) cô gái(女孩) xinh đẹp (美丽的).”；只有表示数量的词语（数词、量词）或指示代词（各、每等）充当定语时，定语排在中心语之前，例如，汉语句子“我吃了一个苹果。”，对应的越南语“Tôi(我) ăn(吃) một quả (一个) táo(苹果).”；（2）越南语与汉语描写性定语的位置完全不同，但定语修饰中心语的顺序(定语与中心语的远近距离)一致，越南语描写性多层定语的结构顺序与汉语呈镜像关系，汉语中描写性定语的顺序是:1-2-3-4-中心语；与之相反，越南语的顺序是：中心语-4-3-2-1。例如，汉语句子“她是我见过的最美丽的女孩。”，对应的越南语“Cô là（她是） cô gái（女孩）xinh đẹp nhất（最美丽的） mà tôi từng thấy（我见过）”；（3）越南语状语成分与汉语大多数情况下是一致的，但汉语常把表示时间的状语放在主语之后,而越南人更习惯把表示时间的状语放在句首，另外,越南语表示时间的状语若是由介词短语充当,其位置常在句末。例如，汉语“他今天没来上课。”，越南语“Ngài không đến lớp học ngày（他没来上课）hôm nay(今天)”；（4）越南语表示处所的状语一般位于谓语动词之后,与汉语不同。例如，汉语“我常常在食堂吃饭。”，越南语“Tôi thường ăn (我常常吃饭) ở quán ăn tự phục vụ (在食堂).”（5）题语一般放在主语前（若出现宾语前置时即被动式表示时，此宾语越南语也称为“题语”）。例如，汉语“他寄信走了。”，越南语为“Thư（信） nó（他）gửi（寄）đi rồi（走了）.”（此句中，信是题语）；有时放在主语后，称为“次题语”，如句子“ông giáo ầy（那个老师），thuốc（烟）không hút（不抽），rượu（酒）không uống（不喝）。”中，“烟”和“酒”就是次题语。

3 汉语—越南语词对齐

词对齐是统计机器翻译中一个非常重要的概念，图 1 给出了一个汉语句子和一个越南语句子词对齐的例子。在这个实例里有 6 个需要对齐的词对：Tôi(我） là(是）Trương(张) thầy(老师) của(的) học trò(学生)。本文中，我们以 P.F.Brown 等人的表示方法为例[10]，那么这个汉语--越南语句对词对齐的关系可以表示成如下形式：(Tôi là học trò của thầy Trương|我(1)是(2)张(6)老师(5)的(4)学生(3))。其中，汉语单词后面的数字表示的是越南语句子中与其对齐的越南语单词的位置。例如，学生(3)表示名词学生与越南语语句子中的第 3 个单词 học trò 对齐。本文

使用开源工具 GIZA++[8]来对汉语—越南语的平行句对进行词对齐处理，得到的词对齐结果准确率为49.32%，所以需要再进行人工调整校对，词对齐的语料都是平行句对校对的时候就是由相关人员进行一一甄别校对的，这里没有做统一的规范，调整的词数量大概是 40 万词，最后得到高质量的词对齐平行句对。GIZA 软件包最早由约翰.霍普金斯大学的机器翻

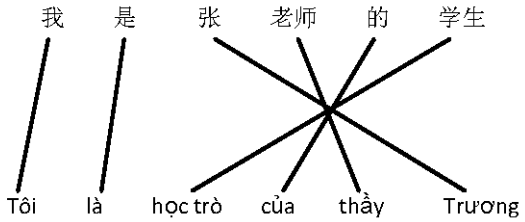

图 1 词对齐的例子

译夏令营实现的，后来，Och 等人对 GIZA 软件包进行了优化处理，称之为 GIZA++。GIZA++实现了 IBM 公司提出的 5 个机器翻译模型，它的主要思想是利用双语平行语料来进行词对齐训练，由句子对的训练得到词语的对齐结果。现今，GIZA++依然是大部分统计机器翻译系统的核心构成部分，在词对齐方面有着广泛的应用。

4 汉语的依存分析

句法分析的任务是根据给定的语法，自动推导出句子的语法结构。目前，在句法分析的研究中主要有短语结构语法和依存语法。短语结构树由终结符、非终结符以及短语标记这三种符号按照特定的语法规则构成。短语结构语法规则，若干终结符构成一个短语，作为非终结符参与下一次归约，直至将整个句子归约为根节点。依存语法认为句子中的述语动词是支配其他成分的中心，而它本身却不受其他任何成分的支配，所有的受支配成分都以某种依存关系从属于其支配者。可以看出，依存语法以其形式简洁、易于标注、便于应用等优点，逐渐成为当今研究人员的研究主题。因此依存语法的研究在许多种语言中均已开展。本文实验中采用了依存语法作为句法分析的语法体系[9]。图 2 为一棵汉语依存句法树，从图中可以看出：依存语法的表示形式简洁，易于理解。依存语法直接表示词语之间的关系，没有额外增加语法符号。所以即使是非专业的人也很能理解该语法形式，这对树库的建设工作十分有利。

汉语和越南语主要的语义关系相似，汉语的依存句法分析是建设越南语依存树库的前提。针对越南语的结构特点和语义关系，同时也为了避免数据稀疏问题，本文定义了如表 1 所示的依存关系集，实验主要基于所定义的 14 种主要依存关系。

5 汉语到越南语句法树的映射

基于前述的汉语—越南语词对齐以及针对汉语的句法分析，接下来要做的就是从汉语到越南语的依存关系映射，即根据汉语依存句法树和汉语—越南语词对齐的关系，进一步生成越南语的依存句法树。对两种语言进行研究发现，虽然越南语子中的词序与汉语句子中的词序不一致，但是依存关系却是一致的，所以可以把汉语句子的依存关系直接映射到越南句子上，具体方法如下例所示：

越南语：cô ấy người chồng côngtác ở Canada（1-1）

汉语：她的先生在加拿大工作。（1-2）

经过词对齐处理之后的结果为：

越南语：cô ấy(1) người chồng(2) côngtác(3) ở(4) Canada(5)（2-1）

汉语：她(1)的先生(2)在(4)加拿大(5)工作(3)（2-2）

以上括号中数字代表其前面的词在本句子中的顺序。对汉语句子进行句法分析得到汉语的句法分析树，如图 3 所示。

接下来我们要做的就是结合越南语言的语法特点，并基于前面的词对齐和汉语的依存句法树来生成越南语句子的依存句法树，如图 4 所示。

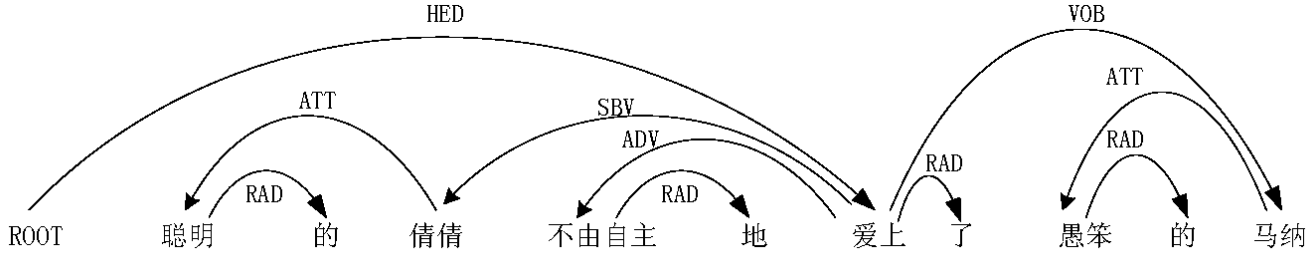

图 2 汉语依存树的结构

表 1 越南语依存关系表

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我给他一个苹果 (我 <-- 给)
动宾关系	VOB	直接宾语，verb-object	我给他一个苹果(给 --> 苹果)
间宾关系	IOB	间接宾语，indirect-object	我给他一个苹果(给 --> 他)
前置宾语	FOB	前置宾语，fronting-object	他任何水果都吃 (水果 <-- 吃)
兼语	DBL	Double	妈妈叫我吃饭 (叫 --> 我)
定中关系	ATT	Attribute	小白杨 (小 <-- 白杨)
状中结构	ADV	Adverbial	十分迅速 (十分 <-- 迅速)
动补结构	CMP	Complement	吃完了饭 (吃 --> 完)
并列关系	COO	Coordinate	大树和小草(大树 --> 小草)
介宾关系	POB	preposition-object	在屋子里 (在 --> 里)
左附加关系	LAD	left adjunct	大树和小草(和 <-- 小草)
右附加关系	RAD	right adjunct	同学们 (同学 --> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	Head	指整个句子的核心

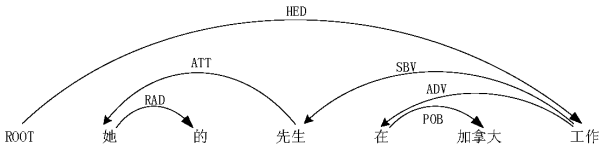

图 3 汉语句子的依存树

从图 4 中可见，尽管汉语句子中的“工作”和越南语句子中的“côngtác”在句子中的前后顺序不一致，但是对依存关系没有影响。经过对越南语和中文语法结构的研究发现两种语言的依存结构几乎存在等价性。所以，可以直接把汉语句子的依存关系直接映射到越南语句子上，来生成越南语的依存句法分析树。然而，由于两种语言的差异性，仍会导致映射存在歧义性，本例中，从式（2-1）和（2-2）句对中的词对齐可以看出，中文句子中的“的”对空了，所以式（2-2）中文句子的“的”相关的依存关系没有映射对象，从图中可以看出越南语句子的依存关系都已经分析出来了，因此这并不影响对越南语句子的分析效果[10]。

越南语中存在一些词对应一个汉语短语的情况，本文总结出一个特殊越南语词典如表 2 所示。字典中，有 132 个越南语词，每一个越南语词都对应着一个汉语短语，除了这些相对特殊的越南语词外，其他的越南语词语与汉语词语的关系基本上都是一一对应的。

实验中，这些特殊越南语词的依存关系我们是根据汉语短语中的核心词来判定的，文中规定：汉语短语的核心词就是依存树中短语部分的根节点。用越南语词“tínöõa”来做一个映射实例，“tínöõa”的汉语意思是“打电话”，如图 5 所示。

表 2 汉语短语——越南语词的对照表

特殊越南语词	汉语短语
giôõn	开玩笑
tí nöõa	再来一点儿
tínöõa	打电话
nghe	听电话
truy tì m	人肉搜索
...	...

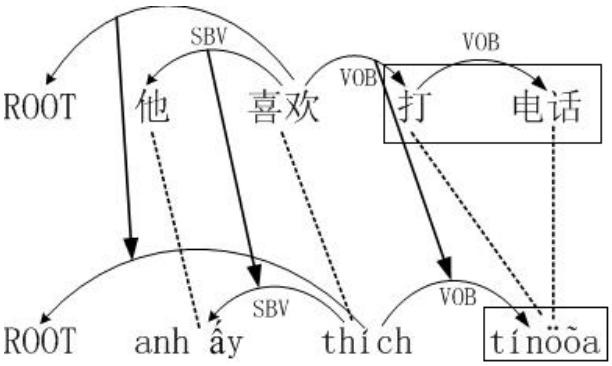

图 5 越南语依存树生成方式二

图 5 中，特殊越南语词“tínöõa”对应汉语短语“打电话”，而特殊越南语词“tínöõa”的依存节点与汉语中“打电话”中的核心词“打”是一致的，“tínöõa”依

存关系与短语核心词“打”也是一致的。经过对汉语句子和越南语句子的对照研究发现，大多数特殊越南语词的依存关系和依存节点与其对应的汉语短语中的核心词在是一致的，可以通过汉语短语中的核心词来确定特殊越南语词的依存节点和依存关系。

6 实验及结果分析

6.1 实验数据

实验数据是来自 7 个新闻网站的国际频道的新闻。这些网站覆盖了各大主流的新闻网站，且包含的新闻覆盖：体育、政治、娱乐、军事等各个方面，因此，保证了实验数据的多样性。

6.2 评价方法

整句依存句法分析评测指标选择：依存弧准确率（Unlabeled Attachment Score,UAS）、标识准确率（Labeled Attachment Score，LAS）和根节点正确率（Root Accuracy，RA），定义如下：

$$UAS = \frac{\text{弧正确的词数}}{\text{所有词数}} \times 100\%$$

$$LAS = \frac{\text{依存弧正确并且依存关系正确词数}}{\text{所有词数}} \times 100\%$$

$$RA = \frac{\text{根正确的句子数}}{\text{句子总数}} \times 100\%$$

6.3 结果分析

采用汉语为中介构建越南语依存树库的方法使用的是 30,000 条汉语—越南语句子对；汉语的依存句法分析是采用哈尔滨工业大学的 LTP 平台 [11]完成的，LTP 工具的标注集我们按照实验的要求和越南语的特点进行了统一的改动；通过汉语—越南语的映射生成 30,000 条越南句子的依存树库。3 万句的语料是第一阶段的语料，随着语料的不断增加，实验也会不断的完善。分别统计数量分别为 10,000、20,000、30,000 条句对的实验结果，如表 3 所示。

表 3 汉语为中介构建越南语依存树库实验结果

语料数（句对）	UAS%	LAS%	RA%
10,000	78.16	74.21/	82.45
20,000	79.36	74.56/	83.56
30,000	79.27	73.89	83.96

同时，本文以 5,000 条人工标注的越南语句子为初始集，用 MaltParser[12]和 MSTParser[13]工具对其进行机器学习建模，生成依存树模型，再用生成的越南语依存树模型对越南语句子进行扩展。实验中扩展了 30,000 句越南语依存树库。这样，我们就有了基于统计机器学习方法生成的依存树库。使其与采用汉语为中介构建的越南语依存树库的实验方法进行比较。实验结果如表 4 所示。

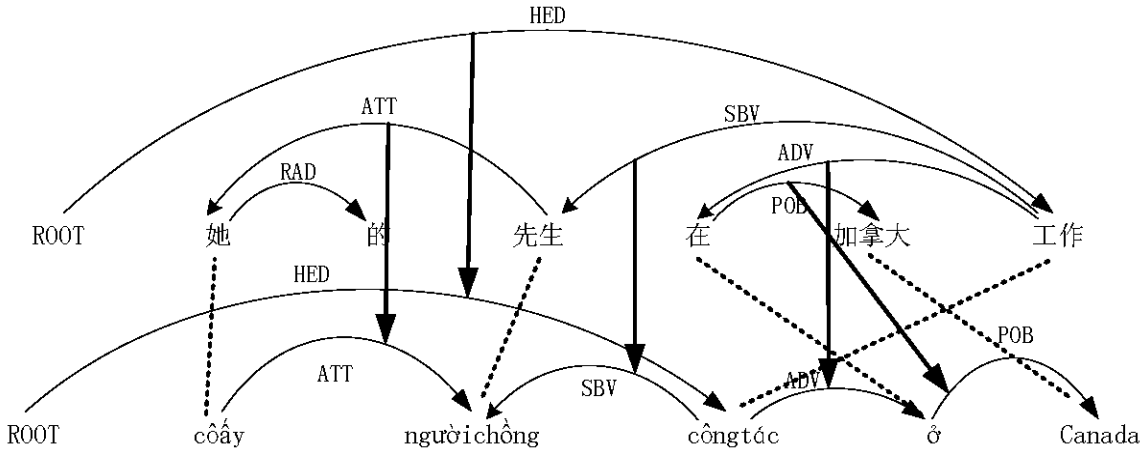

图 4 越南语依存树生成方式一

表 4 其他方法和本文方法的比较

方法	UAS%	LAS%	RA%
MaltParser 构建的越南语依存树库	76.08	71.66	81.79
MSTParser 构建的越南语依存树库	75.03	71.12	80.85
采用汉语为中介构建的越南语依存树库	78.93	74.22	83.32

从表 3 和表 4 中可以看出，在越南语语料相对比

较少的情况下，采用以汉语依存库为基础，基于规则的映射方法所生成的越南语依存树库，准确率相比采用机器学习的方法明显提高。

将5,000句人工标注数据和30,000句利用中间语转化的数据一起训练依存分析模型，然后用来训练新的越南语依存句法树，得到的依存树的准确率会比以5,000句人工标注的数据低一些，而self-training之后得到的依存树的准确率又低一些。这是本文提出的方法得到的越南语依存树库存在一些错误造成的。

分析实验结果，由于越南语言结构在一定程度上和中文语言结构类似，但又具有其特殊的语言特点，因此可以采用以汉语依存库为基础、基于规则的映射方法来生成越南语的依存树库，这样可以避免越南语语料的人工标注过程；在越南语语料相对少的条件下，可以获得比机器学习高的准确率。随着语料的不断增加，机器学习的baseline的准确率也会得到相应的提高。本文对错误实例经过分析发现，本文提出的方法对短句效果好，而长句的处理效果相对较差。这是由于长句句式复杂，且两种语言有很大差异分，还需结合深层次的语言结构分析。还有一部分错误是由中文依存自动分析结果不准确造成的。另外，通过本文方法得到的依存树存在有些词和句中其他词之间不存在任何依存关系的情况，而人工标注的越南语依存树不存在这种情况，这也是由两种语言之间的差异造成的。在下一步的研究中，我们将针对长句依存关系和两种语言之间的差异进行研究，同时会对中文的依存结构进行校正调整，不断提高中文依存树库的准确率，最后得到准确率更高的越南语依存树库。

结束语

本文提出了基于汉-越语言对齐关系的越南语依存树库的构建方法，该方法避免了人工标注越南语依存树库的过程。相对于传统的统计机器学习的方法此方法更加简单，准确率得到了一定程度上的提升。解决了越南语依存树库资源建设困难等问题。下一步，我们将基于不同语言与越南语对齐关系进行越南语依存树库的构建实验，并与基于汉—越语言对齐关系构建的越南语依存树库进行比较分析，最终实现融合多语-越南语的对齐特性来进行越南语依存树库的构建实验。

参 考 文 献

[1] 马金山. 基于统计方法的汉语依存句法分析研究[D], 哈尔滨工业大学,2007

[2] J. Hajic. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank[C]//Issues of Valency and Meaning,1998,106-132

[3] Tracy Holloway King, Richard Crouch, Stefan Riezler,. Mary Dalrymple and Ronald M. Kaplan. The PRAC700 dependency bank[C]//Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora(LINC-03). 2003,1-8

[4] I. Boguslavsky,S. Grigorieva, N. Grigoriev, L. Kreidlin and N. Frid. Dependency treebank for Russian: concept, tools, types of information[C]//The 18th International Conference on Computational Linguistically(COLING),2000,987-991

[5] C. Bosco and V. Lombardo. Dependency and relational structure in treebank annotation.[C]//WorkShop on Recent Advances in Dependency Grammar,2004:1-8

[6] 陈鑫. 基于主动学习的汉语依存树库构建[D], 哈尔滨工业大学,2011

[7] P. T. Nguyen, L. V. Xuan, T. M. H. Nguyen, and P. LeHong. Building a large syntactically-annotated corpus of Vietnamese[C]//Proceeding of the 3th Linguistic Annotation Workshop, ACL-IJCNLP, Singapore, 2009,182-185

[8]SU Xiang,LI Yu-jian.Computational Performance Analysis of GIZA++.[J].COMPUTER ENGINEERING & SCIENCE,2010

ztyu@bit.edu.cn

[9] 车万翔, 张梅山, 刘挺. 基于主动学习的中文依存句法分析[J]. 中文信息学报, 2012,2(6),18-22

[10] Luong Nguyen Thi ,Dalat Univ,Lamdong,Vietnam ,Linh Ha My,Hung Nguyen Viet,Huyen Nguyen Thi Minh,Phuong Le Hong.Building a Treebank for Vietnamese Dependency Parsing[C]//IEEE RIVF International Conference on Computing and Communication Technologies - Research, Innovation, and Vision for the Future (RIVF), NOV 10-13, 2013

[11] http://ir.hit.edu.cn

[12] Joakim Nivre,Johan Hall andJens Nilsson. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing[C]//Proceedings of the fifth international conference on Language Resources and Evaluation, 2006, 2216-2219

[13] R. McDonald, K. Lerman,and F. Pereira. Multilingual Dependency Analysis with a Two-Stage Discriminative Parser[C]// Tenth Conference on Computational Natural Language Learning ,2006, 216-220

作者简介：

李发杰（1986——），男，硕士研究生，主要研究领域为自然语言处理与句法分析。
Email: lfj100120@163.com

郭剑毅（1964——），女，硕士，教授，硕士生导师，主要研究领域为自然语言处理、信息抽取、机器翻译等。
Email: gjade86@hotmail.com

余正涛（1970—），博士，男，博士生导师，教授，主要研究领域自然语言处理、信息检索、机器翻译、机器学习等。
Email: ztyu@bit.edu.cn