文章编号：1003-0077（2018）00-0000-00

基于汉维映射关系构建维吾尔语依存树库

吐尔洪·吾司曼[1,2,3]，杨雅婷[1,2,3]，王磊[1,2,3]，周喜[1,2,3]，程力[1,2,3]

（1.中国科学院 新疆理化技术研究所，新疆 乌鲁木齐 830011；2.中国科学院大学，北京 100049；3.新疆民族语音语言信息处理实验室，新疆 乌鲁木齐 830011）

摘要：该文提出一种基于汉语依存句法信息来构建维维吾尔语依存句法树库的方法。首先对维吾尔语进行形态分析、之后进行汉维词对齐、中文依存分析，然后根据词对齐信息以及汉语依存信息得到维吾尔语依存信息，最终对结果进行优化获得维吾尔语依存句法库。在此基础上训练得到的依存句法分析器在CoNLL 2017 Shared Task 测试集上进行实验，带标记依存正确率LAS(Labeled Attachment Score)和无标记依存正确率UAS(Unlabeled Attachment Score)分别为34.38%和52.53%。

关键词：维吾尔语；依存句法；映射

中图分类号：TP391 **文献标识码：**A

Build Dependency Treebank of Uyghur based on Chinese-Uyghur Mapping

Turghun Osman[1,2,3], YANG Yating[1,2,3], WANG Lei[1,2,3], ZHOU Xi[1,2,3], and CHENG Li[1,2,3]

(1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, Xinjiang 830011, China;
2. University of Chinese Academy of Science, Beijing 100049, China;
3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, Xinjiang 830011, China)

Abstract : This paper presents a novel approach to build Uyghur dependency treebank using dependency information from Chinese-Uyghur bilingual corpus. By analyzing the Uyghur morphology and dependency parsing for Chinese language, we can map Chinese dependency information to Uyghur sentences through word alignment and then further optimize the dependency information for Uyghur language Experimental results in CoNLL 2017 Shared Task testing dataset show that our parsing model can achieve 34.38% and 52.53% improvements on the Labeled Attachment Score(LAS) and Unlabeled Attachment Score(UAS) metrics.

Key words: Uyghur; Dependency Grammar; Mapping

0 引言

句法分析是自然语言处理中的关键技术之一，其基本任务是确定句子的句法结构或句子中词汇之间的依存关系[1]，从而有助于提高人类对自然语言的理解程度，因此句法分析研究具有重要的理论依据和广泛的应用前景。目前句法分析的结果可以用在统计机器翻译[2]、情感分析[3]、问答系统[4]及自动摘要[5]等应用系统中。

依存句法是句法分析的重要分支之一，由于其关注的对象是句中词与词之间的句法功能关系，因此容易表示和理解。早期的依存句法分析主要采用基于规则的方法，有动态规划算法、约束满足的方法以及确定性分析策略等，而随着基于统计的方法在自然语言处理领域中的广泛使用，出

收稿收稿日期：2018-00-00；**定稿日期**：2018-00-00

基金项目：中国科学院“西部之光”基金(2017-XBQNZ-A-005,2017-XBZG-BR-001)；国家千人计划项目(Y32H251201)；国家自然科学基金(U1703133)；新疆自治区重大科技专项(2016A03007-3)

现了生成式依存分析、判别式依存分析和确定性依存分析等一批基于数据驱动的分析方法[6]，并相继构建汉语[7]、俄语[8]、荷兰语[9]、捷克语[10]及土耳其语[11]等语言的依存树库，这些树库所采用的标注规则及标注集互相不兼容，因此后来又提出了跨语种依存句法标注框架 UD (Universal Dependencies)[12]，旨在促进多语种分析器的开发、跨语种学习以及从语言学角度进行分析研究。

维吾尔语是一门相对于汉语、英语而言资源缺乏语言，虽然新疆大学和新疆师范大学在维吾尔语语料库建设方面做了大量的工作[13-15]，不过大部分研究工作集中在词法层面展开，并且这些语料库到目前为止还没有可用的公开版本。维吾尔语依存句法研究是维吾尔语自然语言处理中起步比较晚的研究方向之一，目前面向句法分析的维吾尔语树库建设工作还处在开始阶段，文献[16]开展维吾尔语语义角色标注集的研究；文献[17-18]中讨论如何建立维吾尔语依存关系标注语料库及标注规范等问题，并介绍了维吾尔语依存树库 UDT (Uyghur Dependency Treebank)规范；文献[19]提出了从 UDT 到 UD 的转换方法。

树库是一种属于深度加工的语料库，其中不仅包括词干、词性以及词缀等形态学层面的信息之外，还包含句法结构、句法功能及语义角色等句法层面的信息，是给统计自然语言处理研究提供了数据支撑。虽然通过人工标注构建树库具有规范化、质量高等特点，但是也存在标注难度大、耗时长、成本高等弊端，因此出现了利用已有语言的树库，借助知识转移、词对齐等方法获取第二种语言树库的研究工作[20-21]。

本文借鉴双语之间依存信息的转移思想，利用现有的汉维平行语料库，通过汉维双语句子之间的词对齐关系，把汉语句子的依存关系映射到维吾尔语句子获得维吾尔语依存树库。在此过程中维吾尔语的复杂形态特性导致数据稀疏问题，从而降低对齐精度，最终影响到映射结果的正确性；而汉维语种之间的差异也会影响最终的结果。针对复杂形态导致的数据稀疏问题，本文先做形态分析获取词干，再进行词对齐；而针对语种之间差异造成的问题，则利用语言学规则来降低其对最终结果的影响。

1 研究背景

1.1 维吾尔语依存句法

维吾尔语是一种黏着性语言，其句法具有如下特性：

第一，维吾尔语是采用 S-O-V 结构（主-宾-谓）的语种，采用从右往左的书写方式；

第二，维吾尔语句子的主语可以省略，被省略的主语可以根据谓语的人称词缀判断得到，如“بۈگۈن مەكتەپكە باردىم”(今天我去学校了)，虽然句子中没有出现主语“مەن”（我），但是对谓语进行形态分析之后得到的切分形式为“بار+د+ىم”，其中“ىم”为表示“第一人称单数”的人称词缀，因此我们很容易确定原句的主语是“مەن”；

第三，维吾尔语句子中单词之间的协调由构形词缀来完成的，它们不仅能确保句子的通顺，而且还携带相应的语义信息。维吾尔语具有丰富的形态特性，一般情况下词干由不同的词缀构形之后所表达的意思也发生变化，如词干“ئۆي”（房间）与属格“نىڭ”构形之后表示“ئۆينىڭ”（房间的），与时位格“دە”构形之后表示“ئۆيدە”（在家），与第一人称单数词缀“ۇم”构形之后表示“ئۆيۈم”（我的房间）。在句子中这些构形词缀除了能表达词法层面的意义外，还能表达句法层面的意义，如，主语跟谓语的人称、数等信息必须一致；由宾格构形的单词做宾语；主语必须是主格形式等。除此之外，单词的词性对确定其句法功能也有帮助，如：形容词修饰名词时做定语，修饰动词时做状语；副词修饰动词做状语；名词、代词、形容词、数词或者名词短语可做主语；动词可做句子的中心词等。图 1 显示例句 1:

گۇاڭجۇ بەيیۇن ئايرودۇرۇمى بۈگۈن جەمئىي 436 نۆۋەتچى ئايروپىلاننىڭ ئۇچۇشنى ئەمەلدىن قالدۇردى.

（广州白云机场今天一共取消了 436 个航班。）的依存分析图。

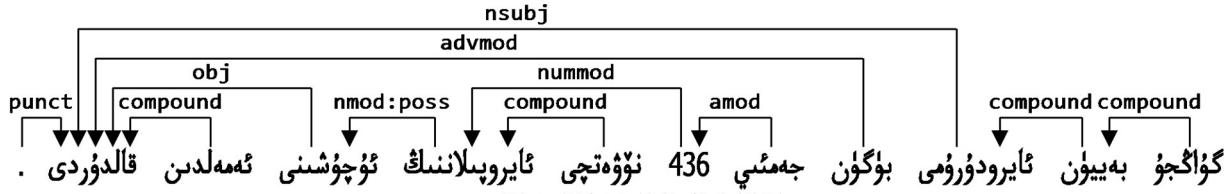


图 1 例句 1 的依存分析图

其中主语“گۇاڭجۇ بەيیۇن ئايرودۇرۇمى”（广州白云机场）依存于谓语“ئەمەلدىن قالدۇردى”（取消了），形成主谓关

系（nsubj）；因为“گۇاڭجۇ بەيىۋن ئايرودۇرۇمى”、“ئەمەلدىن قالدۇردى”和“نۆۋەتچى ئايروپىلان”（航班）都是词组，因此词组中的前一个词依存后一个词形成复合关系（compound）；副词“بۈگۈن”（今天）依存于谓语，形成状中关系（advmod）；形容词“جەمئىي”（一共）修饰数字“436”，形成定中关系（amod）；数字“436”修饰“نۆۋەتچى ئايروپىلان”，构成数量关系（nummod）；词组“نۆۋەتچى ئايروپىلاننىڭ”依存于名词单词“ئۇچۇش”（起飞），形成领属关系（nmod:poss）；宾语“ئۇچۇشنى”（起飞）依存于谓语，形成动宾关系（obj）；标点符号“.”（。）依存于中心词，依存关系为 punct(标点符号)。表 1 给出例句 1 的形态分析和依存分析结果。

表 1 形态和依存分析结果

序号	单词	词性	词缀	词性	位置	关系
1	گۇاڭجۇ	名词	-	-	2	复合
2	بەيىۋن	名词	-	-	3	复合
3	ئايرودۇرۇمى	名词	ى	第三人称	11	主语
4	بۈگۈن	副词	-	-	11	状中
5	جەمئىي	形容词	-	-	6	定中
6	436	数词	-	-	8	数量
7	نۆۋەتچى	形容词	-	-	8	复合
8	ئايروپىلاننىڭ	名词	نىڭ	属格	9	领属
9	ئۇچۇشنى	动词	نى	宾格	11	动宾
10	ئەمەلدىن	名词	دىن	从格	11	复合
11	قالدۇردى	动词	ى	第三人称	0	中心词
12	.	标点符号	-	-	11	标点符号

维吾尔语依存句法除了包括上述例子中的主谓关系、动宾关系、状中关系、定中关系、复合关系、领属关系、标点符号及中心词等依存关系之外，常用的依存关系还有：

（1）并列关系（conj）：句子中并列成分之间的关系由并列关系表示，并列关系中从属成分与支配成分在句子中的作用是平等的，如例句 2：“مەن بازاردىن ئالما، نەشپۈت ۋە بانان سېتىۋالدىم.”（我从市场上买了苹果、香梨和香蕉。）中“ئالما”（苹果）跟“نەشپۈت”（香梨）、“نەشپۈت”跟“بانان”（香蕉）之间形成并列关系。

（2）连词关系（cc）：表示连词与被连接成分之间的依存关系，如例句 2 中的“ۋە”（和）依存于“بانان”形成连词关系。

（3）表系关系（cop）：表示表语和系动词之间的依存关系，如例句 3“ئۇنىڭ كىيىنىشى رەتلىك بولدى.”（他的衣着整齐。）中的“بولدى”（是）依存于“رەتلىك”（整齐），形成表系关系。

(4)同位关系（appos）:同位语之间的关系由同位关系表示，如例句 4“بىز ئۆزىمىز ھەل قىلىلى.”（我们自己解决吧。）中,“بىز”（我们）依存于同位词“ئۆزىمىز”（自己），形成同位关系。

(5)体助关系（aux）:用于表示体助动词与动词之间的依存关系，如例句 5“تېز مېڭىپ باقە.”（走快点。）中的体助动词“باقە”跟“مېڭىپ”（走）之间存在体助关系。

(6)补足关系（ccomp）:维吾尔语中补足关系主要发生在由动词“دە”（说）跟其引用的内容之间，如例句 6“ئۇ ماڭا ياردەم بېرىمەن دېگەن ئىدى.”（他说过要帮助我。）中的“دېگەن”（说过）做支配成分,“بېرىمەن”（给予）做从属成分，形成补足关系。

(7)限定关系（det）:限定词跟名词或者名词短语之间的依存关系表示为限定关系，如“مۇشۇ بىنا”（这栋楼）中代词“مۇشۇ”（这）限定后面的名词“بىنا”（楼），构成限定关系。

(8)指向关系（nmod:cau）：指向关系用来表示动作的去向，如“ئاشخانىغا بارىمەن”（准备去食堂）中单词“ئاشخانىغا”（往食堂）依存于“بارىمەن”（我去），形成指向关系。

通过以上分析可以发现，维吾尔语句子中从属成分的形态信息取决于支配成分的形态信息，反过来，从属成分的形态信息在确定依存关系时起到参考作用，因此形态分析对依存句法分析有一定的帮助作用。

1. 2 汉语和维吾尔语的对比

维吾尔语是黏着性语言，黏着性语言的最大特点是具有丰富的形态特性和音变现象（弱化、脱落及增音），其词干按规律由不同的词缀构形，形成新的词语，而且构形过程中还会发生音变现象。

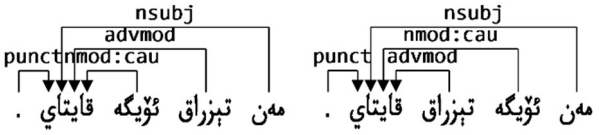
而汉语是分析语（又叫孤立语），其最大的特点是没有形态变化，但有声调，不同的声调，其表达的意思会完全不同。

汉语跟维吾尔语不仅在词法上具有较大的不同，而且句法上也有存在较大的差异：

（1）书写方式不一样，汉语是从左往右书写，而维吾尔语是从右往左书写；汉语句子中词语之间没有分隔符；而维吾尔语句子中的词语必须由空格隔开。

（2）语言结构不一样，汉语句子是 S-V-O（主谓宾）结构，由于没有形态变化，因此由固定的词序来表达意义；维吾尔语是 S-O-V（主宾谓）结构，词序相对而言比较灵活，如例句 7“مەن تېزراق ئۆيگە قايتاي.”中的单词“تېزراق”（尽快）和“ئۆيگە”（往家）互换位置形成例句 8“مەن ئۆيگە تېزراق قايتاي.”，但是他们所表达的意思都是“我要尽快回家”。

虽然汉语和维吾尔语在词法和句法上具有较大的差异，但是依存句法研究的对象是词语跟词语之间的依存关系，和词序、句法结构的关系不大，图 2 分别给出了例句 7 和例句 8 的依存关系图，从图中可以发现，这两个句子的依存关系没有词序的改变而发生变化。



(a)例句 7 (b)例句 8

图 2 依存关系图

并且汉语依存句法跟维吾尔语依存句法中有共同的依存关系，如：主谓关系、动宾关系、领属关系、并列关系、连词关系、状中关系、定中关系等承载主要句法信息的依存关系是一致的。图 3 显示例句 9：

“يېقىندا چاۋشيەن يېرىم ئارلىنىڭ جىددىي ۋەزىيىتى ئۆزلىكسىز ياخشىلاندى.”

的依存关系，图 4 显示对应译文“近期朝鲜半岛紧张局势不断好转。”的依存关系，表 2 显示例句 9 和对应译文之间的词对齐情况。

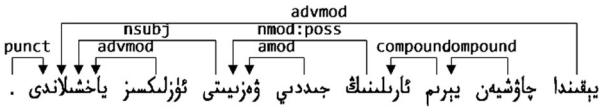


图 3 例句 9 的依存关系图

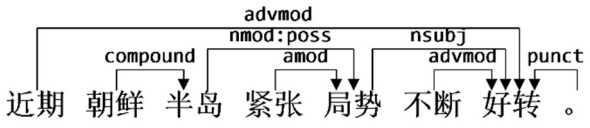


图 4 例句 9 译文的依存关系图

表 2 例句 9 和译文的词对齐矩阵

	近期	朝鲜	半岛	紧张	局势	不断	好转	。
يېقىندا	■							
چاۋشيەن		■						
يېرىم			■					
ئارلىنىڭ			■					
جىددىي				■				
ۋەزىيىتى					■			
ئۆزلىكسىز						■		
ياخشىلاندى							■	
.								■

通过参考表 2 中的词对齐情况，从图 3 和图 4 中可以发现，例句的主谓关系（nsubj）、状中关系（advmod）、领属关系（nmod:poss）、定中关系（amod）以及标点关系（punct）不仅出现在译文的依存关系中，而且依存关系的成分也是相互对齐的。例句中的词组“يېرىم ئارلى”对应译文中“半岛”一词，因此“يېرىم”（半）和“ئارلى”（岛）之间的依存关系为复合关系，而译文中“朝鲜”和“半岛”之间是复合关系，因此例句中的“چاۋشيەن”（朝鲜）、“يېرىم”和“ئارلى”之间依次形成复合关系。

2 基于汉维映射关系的依存信息转移方法

2. 1 算法介绍

本文利用汉维平行语料库，采用依存关系映射的思路，根据汉语句子的依存关系得到维吾尔语句子的依存关系，并借助词法、语法等规则优化最终结果，图 5 是算法流程图。

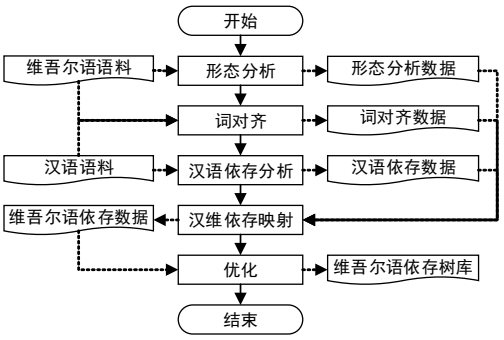


图 5 算法流程图

2.2 维吾尔语形态分析

由于维吾尔语依存句法分析依赖于形态分析，因此，首先对汉维双语语料中的维吾尔语句子进行形态分析，形态分析的内容有词干提取、词缀切分及词性标注。表 3 显示例句 9 的形态分析结果，其中切分形式的第一项为该词的词干。

表 3 例句 9 的形态分析结果

序号	单词	切分形式	词性
1	يېقىندا	يېقىندا	副词
2	چاۋشيەن	چاۋشيەن	名词
3	يېرىم	يېرىم	形容词
4	ئارلىنىڭ	ئارال+ى+نىڭ	名词+第三人称+属格
5	جىددىي	جىددىي	形容词
6	ۋەزىيىتى	ۋەزىيەت+ى	名词+第三人称
7	ئۆزلىكسىز	ئۆزلىكسىز	副词
8	ياخشىلاندى	ياخشىلا+ن+د+ى	动词+反身态+过去式+第三人称
9	.	.	标点符号

2.3 汉维词对齐

本文中汉维之间的词对齐信息是依存关系从汉语映射到维吾尔语的重要依据，因此词对齐结果的好坏直接影响到最终得到的维吾尔语依存树库的质量。

经实验发现，维吾尔语的复杂形态特性导致参与对齐过程的维吾尔语单词数量的增多，对词对齐造成干扰，因此词干提取之后的词对齐结果明显好于没有进行词干提取时的词对齐结果。图6显示例句10“يغىندا يەنە باشقا ئىشلارمۇ مۇھاكىمە قىلىندى .”和译文“会议还讨论了其他事项。”在词干提取之前和之后的对齐的对比结果，其中实线表示标准对齐，点线表示提取之前的对齐结果，虚线表示提取之后的对齐结果，从图中可以发现，词干提取之后的对齐结果更接近于真实的对齐情况。

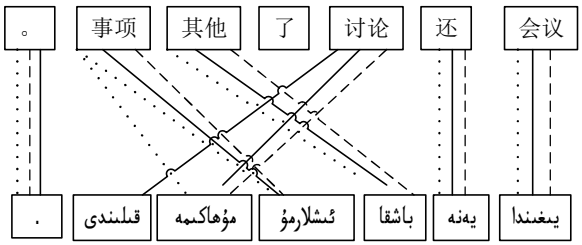


图6 词对齐的对比结果

2.4 依存信息转移

词对齐完成之后，通过对齐信息把汉语句子的依存信息映射到平行的维吾尔语句子，此过程包括预处理和映射。

2.4.1 预处理

预处理要过滤包含过长句子的句对和对齐不合理的句对，其中对齐不合理情况包括如下情况：

1）一个汉语单词对齐过多的维吾尔语单词；

2）一个汉语单词对齐不连续的几个维吾尔语单词，表4给出对齐之后同时出现以上两种不合理情况的对齐例子。

表4 不合理的对齐结果

	她	挨家挨户	宣传	政策	，	做	思想	工作	。
ئۇ	■								
ئۆيمۇ		■							
ئۆي		■							
بېرىپ		■							
سىياسەت				■					
تەشۋىق			■						
قىلىپ			■						
ئىدىيۋى							■		
خىزمەت								■	
ئىشلىدى		■							
.									■

表4中的汉语词语“挨家挨户”不仅对齐到维吾尔语中的四个单词，而且这四个单词不是连续出现的。

2.4.2 依存信息映射

假设依存信息$D_i = \{Rel_i,\ H_i,\ E_i\}$，其中$Rel_i$，$H_i$，$E_i$分别表示第$i$个依存信息的依存关系名称、支配成分以及从属成分，那么一个句子的依存信息可表示为$\boldsymbol{D} = \{D_1, D_2, D_3, \ldots, D_l\}$，其中$l$表示句子中依存关系的数量，因此汉维依存信息的映射函数可由公式（1）表示：

$$\boldsymbol{D_u} = f(\boldsymbol{D_c}, \boldsymbol{M_u}, \boldsymbol{Align_{cu}}) \quad (1)$$

其中$\boldsymbol{D_c}$为汉语句子的依存信息，$\boldsymbol{M_u}$为对应的维吾尔语句子的形态切分信息，$\boldsymbol{Align_{cu}}$为汉维句子的词对齐信息。映射函数f的具体功能是根据词对齐信息$\boldsymbol{Align_{cu}}$，查找依存关系$\boldsymbol{D}_{ci}$中支配成分H_{ci}和从属成分E_{ci}所对应的维吾尔语词做H_{uj}和E_{uj}，并把依存关系Rel_{ci}的值赋予Rel_{uj}，最终形成对应的维吾尔语依存信息$D_{uj} = \{Rel_{uj},\ H_{uj},\ E_{uj}\}$。

由于汉维双语句对在对齐过程中存在一个汉语单词对齐连续几个维吾尔语单词的情况，因此映射之后这些连续的维吾尔语单词当目标依存关系的依存成分，此时这些维吾尔语单词中各词语之间的依存关系暂时未知的。针对具体的对齐情况以及对它们的处理策略如下：

1）H_{uj}, E_{uj}都是单词

这是最为简单的情况，此时直接由Rel_{ci}，H_{uj}及E_{uj}确定映射关系D_{uj}。如：汉语依存关系{‘amod’，‘社会’，‘小康’}中“社会”和“小康”在维吾尔语中分别对应单词“جەمئىيەت”和“ھاللىق”，因此映射过来的依存关系为{‘amod’，‘جەمئىيەت’，‘ھاللىق’}。

2）H_{uj}, E_{uj}中至少有一个是词组

此时假设词组内的单词依次与其后续单词形成未知复合关系（compound:unk），由词组的最后一个元素完成映射操作。如汉语依存关系{‘nsubj’，‘决定’，‘联合国’}中“联合国”在维吾尔语中对应词组“بىرلەشكەن دۆلەتلەر تەشكىلاتى”，“决定”对应“قارار قىلدى”，因此首先形成词组内单词之间的依存关系：

{‘compound:unk’，‘دۆلەتلەر’，‘بىرلەشكەن’}，

{‘compound:unk’，‘تەشكىلاتى’，‘دۆلەتلەر’}，
{‘compound:unk’，‘قىلدى’，‘قارار’}，
最后形成：
{‘nsubj’，‘قىلدى’，‘تەشكىلاتى’}
如果汉语依存关系中的支配单词或者从属单词的词性为名词，其对应维吾尔语词组之间的未知复合关系改成复合关系，因此上述映射关系改为：
{‘compound’，‘دۆلەتلەر’，‘بىرلەشكەن’}，
{‘compound’，‘تەشكىلاتى’，‘دۆلەتلەر’}，
{‘compound:unk’，‘قىلدى’，‘قارار’}，
{‘nsubj’，‘قىلدى’，‘تەشكىلاتى’}
第三条依存关系中的未知复合关系在后续优化部分中进行处理。

2.5 优化

由于通过映射形成的依存信息中存在一些不符合维吾尔语依存句法的依存关系以及还未处理的未知复合关系，因此利用规则来完成进一步优化处理。本文针对以上情况，结合维吾尔语句法给出如下的优化规则：

1） 主语必须以主格形式出现；

2） 主谓关系中的主语和做谓语的动词必须拥有一致的人称和数信息；

3） 带宾格的成分跟动词的依存关系为动宾关系；

4） 名词跟后续的动词之间形成动宾关系；

5） 带属格的成分依存于离它最近，由人称构形的成分形成领属关系，而且支配成分跟从属成分的人称信息必须一致；

6） 体助动词不能当中心词，主动词做中心词，体助动词跟主动词形成体助关系；

7） 副词或副动词跟动词之间的依存关系为状中关系；

8） 形容词和形容词短语跟名词短语形成定中关系；

9） 带向格的成分跟动词形成指向关系；

10）标点符号依存中心词形成标点符号关系；

11） 当名词跟形容词之间形成依存关系，其中形容词为中心词时，其依存关系为主谓关系；

12） 形容词跟动词之间的依存关系为状中关系；

13） 代词修饰名词形成限定关系；

根据以上优化规则（规则 4），之前例子中的未知复合关系{‘compound:unk’，‘قىلدى’，‘قارار’}将变成{‘obj’，‘قىلدى’，‘قارار’}。

3 实验和结论

3.1 实验

本文利用新闻政务领域 60 万句对的汉维平行语料库，首先采用文献[22]的方法对维吾尔语进行形态分析，第二步用 GIZA++[1]进行汉维词对齐操作，第三步用 Stanford Neural Network Dependency Parser[2]进行汉语依存分析，之后用依存转移方法得到维吾尔语依存信息，最后通过基于规则优化得到 3 万句的维吾尔语依存树库。

为了验证所得到的依存树库的质量以及优化规则对树库质量的有效性，本文用 CoNLL 2017 Shared Task[3]提供的训练工具 UDPipe[4]及测试语料 ug-ud-test.conll(100 句)进行模型训练并测试。表 5 给出优化之前的树库训练得到的模型 Model1，优化之后的树库训练得到的模型 Model2 以及 CoNLL 2017 Shared Task 基线系统所提供的模型 Model3 在同样测试集上的得分情况。

表 5 测试结果

模型	UAS	LAS
Model1	46.80%	25.88%
Model2	52.53%	34.38%
Model3	65.43%	40.97%

从表 5 得知，本文训练得到的模型在两个指标上的得分均低于 CoNLL 2017 Shared Task 基线系统所提供的模型，针对得分偏低的情况，本文经分析之后得出以下结论：

1）本文用的 Stanford Neural Network Dependency Parser 在中文树库上 UAS 和 LAS 指标分别得分 83.90%和 82.40%[23]。从图 5 可以发现，实验中各子任务以串行方式执行，它们之间存在一定的错误传播问题，因此分析得到的汉语树库的错误率会影响到最终的维吾尔语依存树库的质量。

2)模型 Model3 是由人工标注的语料库训练得到的，而我们的模型 Model1 和 Model2 训练所用的语料库由算法从汉维语料库中映射而得，因此语料库的质量不如由人工标注而得到的语料库的质量。

3）CoNLL 2017 Shared Task 的测试集是属于文学领域，而本文所用的语料是新闻政务领域，

[1]http://www.statmt.org/moses/giza/GIZA++.html
[2]https://nlp.stanford.edu/software/nndep.shtml
[3]http://universaldependencies.org/conll17/
[4]https://ufal.mff.cuni.cz/udpipe

因此语料领域不同，会对实验结果产生一定的影响。

4）有些句子，在汉语和维吾尔中的表达方式有所不同，因此映射过来的依存关系没能正确表示维吾尔语句子中原有的依存关系。如例句 11“تاشقى سودىدا ئەسلىگە كېلىش خاراكتېرلىك ئېشىش ئەمەلگە ئاشۇرۇلغان.”是被动句子，其主语是“ئېشىش”(增长)，谓语是“ئەمەلگە ئاشۇرۇلغان”（被实现），它们之间存在主谓关系；语料库中对应译文“外贸实现恢复性增长。”的主语是“外贸”，谓语是“实现”，宾语是“增长”，因此译文中“增长”和“实现”之间存在宾动关系，由于映射算法把汉语句子中的动宾关系映射到维吾尔语句子，因此例句 11 的主谓关系被错误的标注为宾动关系。

除此之外，模型 Model2 在两个指标上的得分都比模型 Model1 高，说明实验中所采用的优化规则在处理未知依存关系和不符合维吾尔语依存句法的依存关系时起到作用，因此进一步完善语法规则，可以降低语种差异导致的映射错误问题，从而提高最终得到的维吾尔语依存树库的质量。

3.2 结论

本文介绍了从汉维平行语料库中，借助 GIZA++、Stanford Neural Network Dependency Parser 等工具，利用映射方法获取维吾尔语依存树库的过程。本文的研究成果不仅对人工构建维吾尔语依存书库具有参考意义，而且对研究汉维之间的语义联系也有一定的参考价值。

参考文献

[1] 宗成庆.统计自然语言处理[M].北京：清华大学出版社，2013：179-179.

[2] Kenji Yamada, Kevin Knight. A Syntax-based Statistical Translation Model[C]//Proceedings of 39th Annual Meeting of the Association for Computational Linguistics.2001: 523-530

[3] Dong L, Wei F, Liu S, et al. A statistical parsing framework for sentiment classification[J]. Computational Linguistics, 2015, 41(2): 293-336.

[4] Berant J, Chou A K, Frostig R, et al. Semantic Parsing on Freebase from Question-Answer Pairs[C]//Proceedings of Empirical Methods in Natural Language Processing.2013:1533-1544.

[5] Kikuchi Y, Hirao T, Takamura H, et al. Single Document Summarization based on Nested Tree Structure[C]//meeting of the association for computational linguistics.2014:315-320.

[6] Joakim Nivre. Dependency Grammar and Dependency Parsing[R]

[7] Li Mingqin , Li Juanzi , Dong Zhendong , et al. Building a large Chinese corpus annotated with semantic dependency[C]//Proceedings of the second SIGHAN workshop on Chinese language processing.2003:84-91.

[8] Boguslavsky I M, Grigorieva S, Grigoriev N, et al. Dependency treebank for Russian: concept, tools, types of information[C]//international conference on computational linguistics.2000: 987-991.

[9] Der Beek L V, Bouma G, Malouf R, et al. The Alpino Dependency Treebank[C]//computational linguistics in the netherlands.2002:8-22.

[10] Hajič, J., Pajas, P.: The Prague dependency treebank: annotation structure and support[C]// Proceedings of the IRCS Workshop on Linguistic Databases.2001:105-114.

[11] Oflazer K., Say B., Hakkani-Tür D.Z.,et al. Building a Turkish Treebank. In: Abeillé A. (eds) Treebanks. Text, Speech and Language Technology, vol 20. Springer, Dordrecht.2003

[12] Mcdonald R T, Nivre J, Quirmbachbrundage Y, et al. Universal Dependency Annotation for Multilingual Parsing[C]//meeting of the association for computational linguistics.2013: 92-97.

[13] 玉素甫·艾白都拉.维语句法分析器中的词义排歧问题的研究[J].计算机应用与软件,2002(04):59-62.

[14] 吐尔根·依布拉音,袁保社.新疆少数民族语言文字信息处理研究与应用[J].中文信息学报,2011,25(06):149-156.

[15] 玉素甫·艾白都拉,阿不都热依木·沙力,阿拉帕提古丽.信息处理用维语词汇标注标记集的确定[J].计算机应用,2009,29(07):2006-2008.

[16] 阿里甫·库尔班,吾买尔江·库尔班,房鼎益.维吾尔语框架语义角色标注标记集研究[J].中文信息学报,2013,27(02):41-46.

[17] Samat M, Turgun I, Marhaba E. The Annotation Scheme for Uyghur Dependency Treebank[C]// Asian Language Processing (IALP).2013

[18] Aili M., Xialifu A., Maihefureti, et al. (2016) Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools. In: Murakami Y., Lin D. (eds) Worldwide Language Service Infrastructure. WLSI 2015. Lecture Notes in Computer Science.2016(9442).

[19] Mairehaba A., Weinila M., Tuergen Y. Universal dependencies for Uyghur.[C]// proceedings of WLSI/OIAF4HLT.2016:44-50.

[20] Zeljko Agic, Danijela Merkler, Dasa Berovi. Slovene-Croatian Treebank Transfer Using Bilingual Lexicon Improves Croatian Dependency Parsing.[C]// In Proceedings of IS-LTC.2012:5-9.

[21] Li Y, Guo J, Yu Z, et al. Building vietnamese dependency treebank based on Chinese-Vietnamese bilingual word alignment.[C]//international conference on natural computation.2016: 1330-1335.
[22] Tursun E, Ganguly D, Osman T, et al. A Semisupervised Tag-Transition-Based Markovian Model for Uyghur Morphology Analysis. acm transactions on asian and low resource language information processing, 2016, 16(2):8:1-8:23
[23] Chen D, Manning C D. A Fast and Accurate Dependency Parser using Neural Networks[C]. empirical methods in natural language processing, 2014: 740-750.

杨雅婷（1985-），博士，副研究员，主要研究领域为自然语言处理。
E-mail: yangyt@ms.xjb.ac.cn

吐尔洪·吾司曼（1982-），博士研究生，主要研究领域为自然语言处理。
E-mail：turghun@ms.xjb.ac.cn

程力（1973-），通信作者，博士，研究员，主要研究领域为云计算、生物信息学、高性能计算、软件工程、信息安全。
E-mail: chengli@ms.xjb.ac.cn