蒙古文依存句法分析

苏向东 高光来 闫学亮
(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 近年来,依存句法分析逐渐成为自然语言处理领域中的研究热点。然而,蒙古文的依存句法分析尚未得到足够的重视。基于最大生成树模型在蒙古文依存关系树库TMDT上进行了蒙古文依存句法分析的研究。在简要介绍蒙古文的特点和蒙古文依存关系树库TMDT之后,详细讨论了最大生成树模型。为找到该模型在蒙古文依存句法分析中合适的特征,重点通过实验对8种特征及其组合在句法分析中的性能进行了比较。结果显示,Basic Unigram Features、Basic Bi-gram Features以及C-C sibling Features这3种特征的组合性能最佳。本研究为蒙古文依存句法分析奠定了基础。

关键词 蒙古文,依存句法分析,最大生成树,自然语言处理
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.021

Dependency Parsing for Traditional Mongolian

SU Xiang-dong GAO Guang-lai YAN Xue-liang
(College of Computer Science, Inner Mongolia University, Huhhot 010021, China)

Abstract Dependency parsing has become increasingly popular in natural language processing in recent years. Nevertheless, dependency parsing focused on traditional Mongolian has not attracted much attention. We investigated it with Maximum Spanning Tree (MST) based model on Traditional Mongolian dependency treebank (TMDT). This paper briefly introduced traditional Mongolian along with TMDT, and discussesd the details of MST. Much emphasis was placed on the performance comparisons among eight kinds of features and their combinations in order to find a suitable feature representation. Evaluation result shows that the combination of Basic Unigram Features, Basic Bi-gram Features and C-C Sibling Features obtains the best performance. Our work establishes a baseline for dependency parsing of traditional Mongolian.

Keywords Traditional mongolian, Dependency parsing, Maximum spanning tree, Natural language processing

1 引言

在自然语言处理领域中,句法分析是许多与文档分析相关的应用的关键步骤,诸如:知识抽取、问答系统、自动文摘、情感挖掘等。作为句法分析的一个分支领域,依存句法分析近年来逐渐成为一个研究热点。然而,蒙古文依存句法分析尚未得到足够的重视。主要原因如下:首先,针对蒙古文自然语言处理的研究起步较晚。其次,作为依存句法分析所必需的标注数据集,蒙古文依存关系树库TMDT直到2011年才出现。另外,蒙古语的语法和其他语言有较大不同。这些不同体现在语序、格、时态、语态、领属等方面。这使得蒙古文句法分析中依存关系的确定和分析比较复杂。

本文的出发点是在蒙古文依存关系树库TMDT上,评测最大生成树模型在蒙古文依存句法分析中的性能,并找到该模型在蒙古文依存句法分析中最佳的特征表示。基于最大生成树模型的依存句法分析方法是将寻找最可能的依存关系问题转化为最大生成树的构建问题。本文首先介绍了蒙古文以及实验的数据源蒙古文依存关系树库TMDT,接着详细讨论了最大生成树模型并通过实验重点比较了8种特征及其组合在该模型中的依存句法分析的性能。这8种特征包括:Basic Unigram Features、Basic Bi-gram Features、In Between POS Features、Surrounding Word POS Features、Extended Surrounding Word POS Features、C-P-C Triple Features、C-C Sibling Features以及Word Stem Features。实验表明,基于最大生成树模型的蒙古文依存句法分析方法获得了满意的性能;同时,Basic Unigram Features、Basic Bi-gram Features和C-C Sibling Features这3种特征的组合在句法分析中的性能最佳。

本文第2节回顾了依存句法分析方面的相关工作;第3节描述了蒙古文及其依存关系树库TMDT;第4节讨论了基

到稿日期:2013-05-26 返修日期:2013-07-15 本文受国家自然科学基金资助项目(61263037),内蒙古自然科学基金重大项目(2011ZD11)资助。
苏向东(1984-),男,博士生,助理工程师,主要研究方向为蒙古文信息处理、模式识别与人工智能,E-mail:sxddxs5747@sina.com;高光来(1964-),男,教授,博士生导师,主要研究方向为蒙古文信息处理、模式识别与人工智能、信息检索;闫学亮(1984-),男,硕士,主要研究方向为模式识别与人工智能、信息检索。

于最大生成树模型的依存分析方法；第 5 节列出了最大生成模型中所使用的 8 种特征；第 6 节给出了实验结果，并进行了分析；最后总结全文并指出下一步工作方向。

2 相关工作

下面是与本文研究相关的一些工作。文献[1]将日文依存句法分析转化为寻找最优生成树。文献[2]利用文献[1]的表示方法，提出了针对投影句法结构的基于最大生成树模型的依存分析方法。它使用文献[3]中提出的算法搜索最佳的投影依存关系树。之后，作者又将该方法扩展到非投影句法结构的依存分析[4]，其中用到了文献[5,6]提出的算法。最大生成树模型中参数的学习可以采用最大边界学习算法[7,8]。

依存关系的构建原则和依存关系中父母与子女的识别方法是依存语法的核心。文献[9]讨论了句子在形态、句法和语义 3 个层次上的依存关系。文献[10]详细介绍了依存理论。文献[11]回顾了依存关系的探测方法。蒙古文依存关系树库 TMDT 中定义的依存关系参照了蒙古语语法[12]。

3 蒙古文和蒙古文依存树库

本节介绍了蒙古文的特点和蒙古文依存关系树库 TMDT。后者是本文依存句法分析实验的数据源。

3.1 蒙古文

蒙古语是由维吾尔语衍生出来的一种亚洲语言。它主要在内蒙古以及与内蒙古接壤的一些地区使用。其书写形式是从上到下，由左至右。蒙古文是一种黏着性语言。它通过主干线把字母连接起来构成单词。这些字母分为元音字母和辅音字母。字母的外形根据其在单词中的位置(词首、词中、词尾)的不同而不同。在一些情况下，字母外形的变化是为了与后继字母在外观上协调。

蒙古文单词具有曲折变化和派生的显著特点。这些单词由词干或词干加词缀构成。词缀具有多种语法功能，有的用来体现单复数变化，有的用来体现语态或时态变化。蒙古文名词或相当于名词的词或短语在句中总处于一种格的状态。格成分指示名词或名词短语在句中所充当的成分以及与其他成分之间的关系。以(1)为例，句中单词“ᠶᠢ”是宾格。它出现在“ᠬᠥᠲᠥᠭᠡᠭᠡᠷ”的后面，表明“ᠬᠥᠲᠥᠭᠡᠭᠡᠷ”在句中是宾语。同样单词“ᠨᠢ”是主格，表明“ᠪᠢᠴᠢᠭᠴᠢᠨ”是句中的主语。例(1)同样验证了蒙古文的语序是 SOV 结构，即主语＋宾语＋谓语。在蒙古文依存句法分析中，不考虑单词内部的关系，但要分析格成分与其他成分之间的依存关系。

(1) ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠨᠢ ᠪᠢᠴᠢᠭᠴᠢᠨ ᠬᠡᠮᠵᠢᠶᠡᠯ ᠶᠢ ᠪᠢᠴᠢᠵᠤ ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠪᠠᠶᠢᠨ ᠪᠠᠶᠢᠰᠤ ..

主语　　宾语　　谓语

3.2 蒙古文依存关系树库 TMDT

蒙古文依存关系树库 TMDT 是一个两层的依存关系树库，包括形态层和分析层。形态层标注了语法单元(单词)的词性和词干信息，分析层标注了单词之间的依存关系。图 1 是“ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠨᠢ ᠪᠢᠴᠢᠭᠴᠢᠨ ᠬᠡᠮᠵᠢᠶᠡᠯ ᠶᠢ ᠪᠢᠴᠢᠵᠤ ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠪᠠᠶᠢᠨ ᠪᠠᠶᠢᠰᠤ ..”的标注结果。

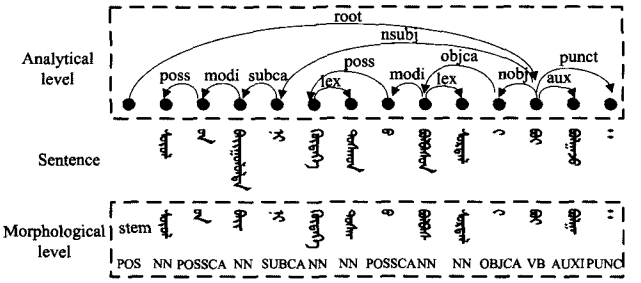

图 1 句子“ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠨᠢ ᠪᠢᠴᠢᠭᠴᠢᠨ ᠬᠡᠮᠵᠢᠶᠡᠯ ᠶᠢ ᠪᠢᠴᠢᠵᠤ ᠬᠥᠲᠥᠭᠡᠭᠡᠷ ᠪᠠᠶᠢᠨ ᠪᠠᠶᠢᠰᠤ ..”标注结果

蒙古文依存关系树库 TMDT 标注的语料来自《内蒙古日报》。该树库共包含 400 个标注的句子，13028 单词，有 4460 个不同的单词和 1548 个不同的词干。在这些句子中，最短的句子有 9 个单词，最长的句子有 54 个单词。平均长度为 32.57个字。其中，简单句占 27%，从句占 73%。

4 最大生成树模型

依存语法认为句中单词与单词之间存在二元的非对称的关联关系。这些关联关系称为依存关系。依存关系中处于从属地位的单词称为子女，处于支配地位的单词称为父母。句中每个单词有且只有一个父母。句中的依存关系不允许环的出现。句子的单词以及依存关系共同构成了句子的句法结构。句法结构可以以有向图的形式表示出来。句中每个单词对应有向图中的一个节点，每个依存关系对应一条从父节点到子节点的有向边。这样的有向图同时也是一棵树，称为依存关系树。

对句子 x 而言，定义它的依存关系树为 $T_x=(V_x,E_x)$，其中

$V_x=\{x_i \mid i\in[1,n], x_i$ 是句子 x 的第 i 个单词$\}$

$E_x=\{(i,j) \mid i\neq j, i,j\in[1,n], x_i$ 在句子 x 中是 x_j 的父节点$\}$

这里，n 是句子 x 的长度；V_x 是句子全部单词的集合；E_x 是句子 x 的依存关系的集合。同时，为句子 x 定义一个有向完全图 $G_x=(V_x,E)$，其中

$V_x=\{x_i \mid i\in[1,n], x_i$ 是句子 x 的第 i 个单词$\}$

$E=\{(i,j) \mid i\neq j, i,k\in[1,n]\}$

显然，T_x 是 G_x 的子图。从自底向上的观点看，依存分析就是为每个单词选择一个父节点来构建依存关系树。即从 G_x 中选择依存关系对应的边来构建 T_x。在依存关系的选择(等价于父节点的选择)过程中需要一个量化的指标。为解决这个问题，定义分数 $s(i,j)$ 来表示单词 x_i 作为单词 x_j 的父节点的概率。需要注意的是，由于父节点的选择比较的是 $s(i,j)$ 的相对大小，因此 $s(i,j)$ 不必一定小于 1 且总和等于 1。前面定义了单一依存关系的分数 $s(i,j)$，那么依存关系树 T_x 总的分数 $s(x,T_x)$ 就是：

$$s(x,T_x)=\sum_{(i,j)\in T_x} s(i,j) \tag{1}$$

因为句子 x 中每个单词 x_j 都是选择对应依存关系分数 $s(i,j)$ 最大的单词 x_i 作为父节点，所以构建的依存关系树 T_x 的分数 $s(x,T_x)$ 一定是句子 x 所有可能的依存关系树中分数最大的一个。根据生成树的定义可知，依存关系树也是生成树。因此，我们可以采用最大生成树模型进行依存句法分析。表 1 列出了本文实验所使用的最大生成树求解算法，即 Chu-Liu-Edmonds 算法[5,6]。算法伪代码中 Contract 和 Expand 操作的具体步骤见参考文献[13]。

表 1　Chu-Liu-Edmonds 算法

Chu-liu-Edmonds(G_x, s)
1. 初始化 G=(V,E), V=V_x, E=∅
2. 如果 G 是生成树，返回 G
3. 否则为 V 中每个顶点贪婪查找分数最高的入边，加入 E 中
4. 如果在 G 中出现圆 C，对 G 执行 Contract 操作变为 G_c， 令 y=Chu-liu-Edmonds(G_c, s)，利用 C 对 y 执行 Expand 操作，返回 y′
5. 否则返回第 2 步

现在的问题是如何定义分数 $s(i,j)$。因为分数 $s(i,j)$ 决定了单词 x_i 和单词 x_j 是否被认定存在依存关系且在依存关系中单词 x_i 是单词 x_j 的父节点。句子中依存关系主要与(1)句子结构、(2)单词 x_i 和单词 x_j 本身以及(3)单词 x_i 和单词 x_j 的语境信息有关。因此，我们给出分数 $s(i,j)$ 的定义如下：

$$s(i,j)=\boldsymbol{w}\cdot\boldsymbol{f}(i,j) \tag{2}$$

其中，$\boldsymbol{f}(i,j)$ 是反映上面 3 个因素的单词 x_i 和 x_j 单词的各自的特征和组合的特征；$\boldsymbol{w}$ 是反映特征权重的系数向量。本文第 5 节给出了实验中使用的 8 种特征的具体定义。权重系数向量 $\boldsymbol{w}$ 在句法分析器的训练阶段利用边界最大化算法[7,8]通过迭代学习获得。

5　特征

本文基于最大生成树模型对蒙古文进行依存句法分析。模型中的特征直接影响句法分析的性能。为了找到合适的特征，实验中对 8 种特征及其组合的性能进行了比较和分析。图 2 列出了实验中所用的 8 种特征。

Basic Unigram Features
p-word
p-pos
p-word, p-pos
c-word
c-pos
c-word, c-pos

C-C Sibling Features
c1-word, c2-word
c1-word, c2-pos
c1-pos, c2-word
c1-pos, c2-pos

Basic Bi-gram Features
p-word, c-word
p-pos, c-pos
p-word, c-pos
p-pos, c-word
p-word, p-pos, c-word
p-word, p-pos, c-pos
p-word, c-word, c-pos
p-pos, c-word, c-pos
p-word, p-pos, c-word, c-pos

C-P-C Triple Features
p-pos, c1-pos, c2-pos

In between POS Features
p-pos, b-pos, c-pos

Surrounding Word POS Features
p-pos-1, p-pos, c-pos-1, c-pos
p-pos-1, p-pos, c-pos, c-pos+1
p-pos, p-pos+1, c-pos-1, c-pos
p-pos, p-pos+1, c-pos, c-pos+1

Extended Surrounding Word POS Features
p-pos-1, p-pos, p-pos+1, c-pos-1, c-pos, c-pos+1

Word Stem Features
p-stem, p-pos, c-stem, c-pos

p-word：依存关系树中的父节点对应的单词，c-word：子节点对应的单词，p-pos：父节点对应的单词的词性，c-pos：子节点对应的单词的词性，p-pos+1：父节点右边单词的词性，p-pos-1：父节点左边单词的词性，c-pos+1：子节点右边单词的词性，c-pos-1：子节点左边单词的词性，b-pos：父、子节点中间单词的词性

图 2　依存句法分析中全部特征

Basic Unigram Features 和 Basic Bi-gram Features 代表依存关系中父子节点对应的单词本身和单词的词性；In between POS Features、Surrounding Word Pos Features 和 Extended Surrounding Word Pos Features 代表依存关系中父子节点对应的单词的语境信息；C-C Sibling Features 和 C-P-C Sibling Features 代表依存关系中父节点以及同一父节点的相邻子女对应的单词的信息；Word Stem Features 代表的是依存关系中父子节点对应单词的词干信息。

6　实验

本文的出发点是测试最大生成树模型在蒙古文依存句法分析中的性能，并找出适合蒙古文依存句法分析的最优特征。本节给出了实验结果并进行了分析。

6.1　实验结果

在蒙古文依存句法分析中，本文采取了 5 组交叉验证的测试方法。评价句法分析性能的 4 个指标分别为：根精度(RA)、带标记依存关系精度(LAS)、完全匹配精度(CM)和无标记依存关系精度(DA)。这 4 个指标的定义如下：

$$RA=\frac{\text{根被正确识别的句子的数目}}{\text{全部句子的数目}} \tag{3}$$

$$LAS=\frac{\text{正确标注了父节点和依存类别的单词的数目}}{\text{单词的数目}} \tag{4}$$

$$CM=\frac{\text{依存分析完全正确的句子的数目}}{\text{全部句子的数目}} \tag{5}$$

$$DA=\frac{\text{正确标注了父节点的单词的数目}}{\text{所有单词的数目}} \tag{6}$$

为了表示方便，我们使用单个字母来代表每个特征。其中：

a：Basic Unigram Features

b：Basic Bi-gram Features

c：In between POS Features

d：Surrounding Word POS Features

e：Extended Surrounding Word POS Features

f：C-C Sibling Features

g：C-P-C Triple Features

h：Word Stem Features

句法分析实验分为带类别标注和无类别标注两个子类。带类别标注的子类的评测指标为 RA、LAS 和 CM；无类别标注的子类的评测指标为 RA、DA 和 CM。根据特征所表示的信息的不同，进一步将带类别标注和无类别标注句法分析实验各分为 5 组，如表 2 所列：1—3，4—10，11—13，14 和 15—17。

表 2　8 种特征及其组合在蒙古文依存句法分析中的性能

Seq.	Features	Unlabeled			Labeled		
		RA	DA	CM	RA	LAS	CM
1	a	73.0%	83.1%	19.0%	72.0%	80.4%	13.0%
2	b	72.0%	81.6%	24.0%	71.0%	79.5%	17.0%
3	a+b	74.0%	83.4%	28.0%	73.0%	81.0%	20.0%
4	a+b+c	67.0%	81.7%	21.0%	66.0%	79.2%	13.0%
5	a+b+d	68.0%	**84.2%**	27.0%	67.0%	**82.0%**	19.0%
6	a+b+e	67.0%	82.2%	**28.0%**	66.0%	80.1%	**20.0%**
7	a+b+c+d	70.0%	82.5%	21.0%	69.0%	80.0%	13.0%
8	a+b+c+e	70.0%	82.2%	21.0%	69.0%	79.8%	14.0%
9	a+b+d+e	70.0%	83.1%	26.0%	69.0%	80.9%	19.0%
10	a+b+c+d+e	73.0%	82.4%	20.0%	73.0%	79.8%	14.0%
11	a+b+f	**74.0%**	**85.0%**	**28.0%**	**73.0%**	**82.6%**	19.0%
12	a+b+g	73.0%	**84.4%**	27.0%	**73.0%**	**82.1%**	19.0%
13	a+b+f+g	70.0%	**84.2%**	**29.0%**	69.0%	**81.8%**	**20.0%**
14	a+b+h	66.0%	82.4%	**28.0%**	65.0%	80.3%	**22.0%**
15	c+d+e+f+g+h	72.0%	82.5%	27.0%	71.0%	79.8%	16.0%
16	a+b+d+f+h	67.0%	**83.8%**	**28.0%**	66.0%	**81.5%**	**20.0%**
17	a+b+c+d+e+f+g+h	**77.0%**	83.2%	23.0%	**76.0%**	81.0%	18.0%

6.2　分析

表 2 列出了蒙古文依存句法分析的实验结果。我们把第 3 组实验的结果作为基准，把性能相同或优于它的结果用粗体字来显示。在第 11 组的实验中，DA 和 LAS 达到了最大值，分别为 85.0%和 82.6%，其他 4 项指标接近于最优性能。因此，Basic Unigram Features、Basic Bi-gram Features 和 C-C sibling Features 这 3 种特征的组合在蒙古文依存句法分析中整体性能最佳。

4 个度量指标的一致性也可以从表 2 观察到。DA 与 LAS 比较一致，而 DA 与 RA、CM 这两者不完全一致。也就是说，DA 和 LAS 的上升和下降是同步的，而 DA 与 RA、CM 的升降是不同步的。比较第 3 组和第 13 组实验可以看到，DA 增大了而 RA 减小了。在第 17 组(所有特征的组合)带类别标注和无类别标注的实验中 RA 均取得最大值，而 CM 并未取得最大值。部分原因是蒙古文语序为 SOV 结构，句中非动词类单词可以作谓语，这很容易与其他一些依存关系混淆。

相比其他组实验，第 14 组带类别标注实验中 CM 达到最大值。但这并不能证明词干特点有助于提高蒙古文依存句法分析的性能。因为相比于第 3 组实验，它的其他指标 RA、DA 和 LAS 均有所下降。在第 13 组无类别标记实验中 CM 达到最大值。

Basic Unigram Features 和 Basic Bi-gram Features 是所有特征中的最重要的两种特征。首先，它们的组合(第 3 组实验)性能优于所有其他特征的组合(第 15 组实验)。其次，它们组合的性能接近于最优性能。

从第 4 组实验到第 10 组实验我们观察到，增加上下文特征并未能明显提升依存句法分析的性能，反而在一些情况下降低了性能。这说明蒙古文依存句法分析中父子节点对应的单词的语境信息对依存关系的判定作用并不明显。这也证实了对基于最大生成树模型的句法分析器而言，特征的选择非常重要。

结束语　本文基于最大生成树模型在蒙古文依存树库 TMDT 上进行了蒙古文依存句法分析的研究。文中详细讨论了基于最大生成树模型的依存句法分析方法，重点通过实验比较了 8 种特征及其组合在蒙古文依存句法分析中的性能，并进行了相应的分析。实验结果显示，Basic Unigram Features、Basic Bi-gram Features 和 C-C sibling Features 这 3 种特征的组合在蒙古文依存句法分析中的性能最佳。无标记依存关系精度 DA 和带标记依存关系精度 LAS 分别达到了 85.0%和 82.6%。考虑到蒙古语语法与其他语言的显著差异以及训练数据的有限性，这个结果是令人满意的。

下一步，我们将尝试在句法分析器中加入蒙古语特性知识，用以提高依存句法分析的精度。

参 考 文 献

[1] Hideki H. Semantic Dependency Analysis Method for Japanese Based on Optimum Tree Search Algorithm[J]. Transactions of Information Processing Society of Japan, 2002, 43(3): 696-707

[2] Mcdonald R, Crammer K, Pereira F. Online Large-Margin Training of Dependency Parsers[C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005: 91-99

[3] Eisner J M. Three New Probabilistic Models for Dependency Parsing: an Exploration[C]// Proceedings of the 16th Conference on Computational Linguistics. 1996: 340-345

[4] Mcdonald R, Pereira F, Ribarov K, et al. Non-projective Dependency Parsing Using Spanning Tree Algorithms[C]// Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005: 523-530

[5] Chu Y-J, Liu T-H. On the Shortest Arborescence of a Directed Graph[J]. Science Sinica, 1965, 14: 1396-1400

[6] Jack E. Optimum Branchings[J]. Journal of Research of the National Bureau of Standards, 1967, 71B: 233-240

[7] Crammer K, Singer Y. Ultraconservative Online Algorithms for Multiclass Problems[J]. Journal of Machine Learning Research, 2003, 3: 951-991

[8] Crammer K, Dekel O, Shalev-Shwartz S, et al. Online Passive-Aggressive Algorithms[C]// Procings of the Sixteenth Annual Conference on Neural Information Processing Systems (NIPS). 2003

[9] Melčuk I A. Levels of Dependency in Linguistic Description: Concepts and Problems[J]. Dependency and Valency, 2003(1): 188-230

[10] Hudson R. An Introduction to Word Grammar[M]. Cambridge: Cambridge University Press, 2010

[11] Nivre J. Dependency Grammar and Dependency Parsing [R]. School of Mathematics and Systems Engineering, Växjö University, 2005

[12] 清格尔泰. 蒙古语语法[M]. 呼和浩特：内蒙古人民出版社，1992

[13] Georgiadis L. Arborescence Optimization Problems Solvable by Edmonds' Algorithm[J]. Theor. Comput. Sci., 2003, 301(1-3): 427-437