北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.021

基于最大熵的汉语篇章结构自动分析方法

涂眉† 周玉 宗成庆

中国科学院自动化研究所模式识别国家重点实验室, 北京 100190; † 通信作者, E-mail: mtu@nlpr.ia.ac.cn

摘要 在标有复句逻辑语义关系的清华汉语树库上, 研究汉语篇章语义片段自动切分以及篇章关系的自动标注方法。比较了不同序列标注模型对汉语篇章语义单元切分的性能, 并提出基于最大熵模型的汉语篇章结构分析方法。实验结果表明, 篇章语义单元自动切分的 F 值能达到 89.1%, 当篇章语义结构树的高度不超过 6 层时, 篇章语义关系标注的 F 值为 63%。

关键词 语义片段自动切分; 篇章结构分析; 逻辑语义关系; 树库

中图分类号 TP391

Automatically Parsing Chinese Discourse Based on Maximum Entropy

TU Mei†, ZHOU Yu, ZONG Chengqing

National Laboratory of Pattern Recongnition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190;
† orresponding author, E-mail: mtu@nlpr.ia.ac.cn

Abstract The authors focus on how to segment semantic units in Chinese discourse and how to label relations among semantic units automatically. During the parsing process, several sequence labelling methods are compared for discourse segmentation, while a maximum entropy-based training and decoding algorithm is specially proposed.Experiments are done based on Tsinghua Chinese Treebank which is annotated with logical and semantic relations at complex-sentence level. Experimental results show that F-score of discourse segmentation reaches 89.1%. When parsing discourses with no more than 6 relations included, the labeling F-score can achieve 63%.

Key words automatic discourse segmentation;discourse structure parsing; Chinese logical and semantic relation; Tsinghua Chinese Treebank

篇章分析是自然语言领域备受关注的一项研究内容。篇章分析是指对篇章的结构和逻辑语义关系进行分析, 以期获得对整个篇章语义层面的理解[1]。例如对于下面一个句子构成的简单篇章, 通过分析, 能得到如图 1 所示的篇章单元和关联关系。图中包括两层篇章结构, 跨度 1 的片段与跨度

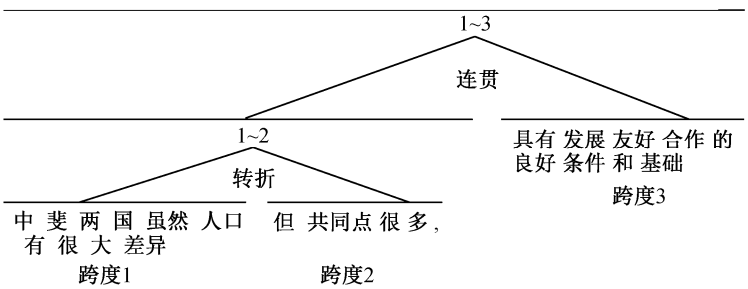

图 1 汉语中逻辑关系的例子

Fig.1 An example of Chinese rhetorical relationship

国家自然科学基金(61003160)、863 计划(2011AA01A207)和中国科学院西部行动计划项目(KGZD-EW-501)资助
收稿日期: 2013–06–15; 修回日期: 2013–09–25; 网络出版时间: 2013-11-08 09:26

2 的片段之间形成转折关系, 然后它们连接起来再与跨度 3 形成连贯关系。

最近 20 年, 随着互联网和计算机技术的高速发展, 依赖于篇章分析和理解的应用不断涌现, 如基于篇章的自动文摘、自动问答以及自动翻译等, 日益凸显出研究篇章分析的重要性和急迫性。以汉英机器翻译为例, 输入图 1 中的汉语文本, 由 Google 在线翻译系统得到的翻译结果为: 1) Although thepopulation ofFijiare very differentbetween the two countries; 2) but have much in common; 3) with good conditions forthe development of friendlycooperationand infrastructure。

由这个翻译例子可知, 由于目前的机器翻译系统无法从篇章的层面获取源语言的整体篇章结构以及篇章中上下文的指代信息, 从而很容易造成翻译结果出现衔接不自然和缺失主语等问题, 如上述的第 2 和 3 句缺少主语, 使得翻译不完整。也就是说, 如果翻译系统不能够正确地分析篇章中各语义单元的逻辑关系和组织方式, 就很难获得语义表达完整和连贯的翻译结果。

尽管有关篇章自动分析的研究工作已经开展了很多年, 涌现出了很多理论方法和技术[1−8], 但是这些工作绝大多数都是以英文为研究对象展开的, 对于汉语篇章的自动分析很少涉及, 到目前仍是一个相对空白的研究难题。

其难点之一: 与传统的句法分析相比, 篇章分析虽然同样是输出层次化的树结构, 但是构成篇章结构树的篇章单元边界的识别比构成句法树的短语边界的识别更加困难, 而且句法树中的节点标识(NP, VP 等)可以用来对整棵树进行 PCFG 推导, 但是篇章树的标识(转折、并列等)并不是某个跨度代表, 它代表的是相关联的跨度之间的关系, 因此 PCFG 不适用于篇章结构的推导, 篇章结构与篇章关系往往需要分开建模; 与英语相比, 汉语篇章单元之间的关系常常没有显式的连接词表示, 获取篇章单元之间的逻辑关系更加复杂, 更难抽取泛化性强的特征, 数据稀疏问题也更加严重。另外, 汉语在遣词造句、谋篇构段上与英语差异显著, 使得两者在标注体系上也有较大差别, 因此亦不能完全照搬英语的标注体系和分析方法。

汉语篇章分析成为难点的另一个重要的原因是缺乏大规模篇章级别的标注语料。乐明[9]提出并标注了基于修辞结构理论的汉语语料, 但是规模较小, 暂未公开。Zhou 等[10]提出基于 Penn 篇章树库的中文篇章树库标注体系, 语料仍在建设中。Song 等[11]提出标注大规模的篇章话题结构的语料, 主要应用于话题结构的识别, 而非逻辑语义结构。清华汉语树库(TCT)[12]标注了复句间的逻辑语义关系, 为我们提供了良好的句子级别的篇章关系资源。

为了充分发挥 TCT 标注语料的作用, 同时又尽量不受句子级别标注的限制, 本文从篇章的角度, 设计了不受限于标注语料的篇章分析方法。通过序列标注分类器将篇章分割成篇章单元后, 我们采用一种基于最大熵的篇章结构和关系标注方法。实验结果表明, 篇章语义单元切分的 F 值最高达到 89.1%。篇章关系分析部分的实验表明, 当篇章树结构的高度不超过 6 层的复句占整个测试集的 93%, 其 F 值能达到 63%。

1 相关工作

有关篇章分析的工作大多是以英文为研究对象, 特别是 RST-DT(RST Discourse Treebank)语料与宾州篇章树库(PDTB)语料发布以后, 带动了大批研究英文篇章分析的工作。

Duverle 等[7]采用多层支持向量机, 从底向上生成修辞结构树。借助此方法, Hernault 等[8]发布了“HILDA”篇章分析器, 这是 RST 框架下目前表现最好的英文篇章分析器, 但是鉴于目前没有公开的 RST 汉语语料, 无法用于汉语篇章分析。另外, 该工作在建树时采用的是局部最优算法, 本文受其启发提出基于最大熵的篇章结构分析算法, 并采用类似 CYK 的全局优化算法进行精确推导。英语的句法和句型比较固定, 而汉语太灵活, 往往需要结合上下文, 才能正确分析篇章结构。特别是没有明显连接词的上下文, 若要正确理解, 必须得全局考虑。因此, 本文的方法在数学上保证生成的树整体得分最高, 并且融入汉语本身的语言特点。

除此之外, 基于宾州篇章树库(PDTB)[13]进行英文篇章分析的研究也逐渐增多[14-16], 其主要是研究如何识别给定的两个相邻的片段之间的篇章关系, 但不会最终形成树型结构, 从而无法把握篇章整体结构。

汉语篇章自动分析的研究工作比较有限, Li 等[17]在宾州汉语树库上人工标注了一些基于 RST 的篇章切分语料, 探讨标点符号在哪些情况下可以作为篇章分隔符。姚双云等[18]提出自动识别复句

中汉语连贯关系的方法, 探讨如何利用关联词语以及其搭配模式自动标注连贯关系, 但这种方法只适用于句子中有关联词语的情况。基于清华汉语树库进行复句分析的研究[19−20]目前仍限于复句级别上, 未涉及篇章语义单元的识别等与篇章理解有关的任务。就目前所掌握的汉语篇章分析的研究动态来看, 本文是首次尝试将汉语篇章语义单元切割和汉语篇章关系自动标注这两者结合起来, 进行完整汉语篇章分析的研究。

2 基于序列标注的汉语篇章单元切分方法

与基于 RST 的英文篇章分析过程类似, 我们的基本思路是首先采用序列标注的方法[21]将汉语篇章切分成篇章单元, 然后用一种基于最大熵的方法建立篇章单元之间的层次结构, 并标注跨度之间的篇章关系。

著名的篇章分析理论“修辞结构理论(Rhetorical Structure Theory)”认为有独立语义以及独立功能的最小片段可以作为基本篇章单元(elementary discourse unit, 简称 edu)。图 1 中叶子节点可以看成是基本篇章单元。对于汉语语义片段的边界来说, 标点是一个相对较好的示性符号或特征, 因此针对标点符号作为篇章单元的边界开展了很多相关研究工作[17, 21]。但是对于很多结构复杂的篇章, 特别是汉语篇章, edu 的形式并不固定, 它可能是一个复句中的小句, 也可能是小句中的一个语义片段, 甚至可能包含不只一个小句。另外汉语中标点符号的使用方法多种多样, 要制定一套基于标点符号的切分规则并不容易。基于以上考虑, 本文以大规模标注树库为指导, 通过序列标注的方式来学习和划分篇章单元的边界。

2.1 篇章单元的表示

沿用文献[21]的表示方式, 每一个篇章单元可以看成是由起始单词和非起始单词连接而成, 起始词用 B 表示, 代表新的篇章单元的开始, 非起始词用 C 表示, 代表当前词仍属于当前篇章单元:

……人口 有 很大 差异 但 共同点 很 多……

……C C C C B C C C……

于是单元切分问题就归约成一个序列标记问题。训练时, 每个词被表示为一个 2 元组<标签, 特征向量>, 比如“但”可以表示为<B, (f_1, f_2, …)>, 解码时通过抽取每个词的特征来预测对应的标签。建立序列标注模型已有大量的相关研究工作, 最著名的模型有最大熵模型、支持向量机模型和条件随机场模型。我们分别用这几种模型以及相应的特征对边界划分问题进行建模。

2.2 篇章单元切分特征

在序列标记问题中, 最重要的任务之一就是选择合适的特征, 使其具有良好的泛化能力和表征能力。在本文中, 受文献[4]和[21]的启发, 我们综合考虑词汇化的特征和句法特征, 下面对其进行详细介绍。

1) 单词和词性(f_1, f_2), 例如 f_1=但, f_2=AD。

2) 短语结构句法树为 t_s, s 中的单词 w 在 t_s 中对应节点的父节点为 N_w, 若 N_w 不是其父节点的唯一子节点, 则考虑 N_w 的句法标记和中心词(lexical head)作为特征 f_3 和 f_4; 若 N_w 的父节点只有唯一子节点, 则从 N_w 向上寻找最高的只有唯一孩子的节点 N_w*, 选择它的句法标签和中心词作为特征 f_3 和 f_4。本文中, 我们称最后得到的父节点 N_w*为最高同源父节点, 如图 2 阴影所示, 图中”条件”的 N_w*为 NP(条件), 所以“条件”的 f_3=NP, f_4=条件。

3) N_w*的父节点的句法标记和中心词(f_5, f_6) 。

4) N_w*的左兄弟节点的句法标记和中心词(f_7,f_8) 。

5) N_w*的右兄弟节点的句法标记和中心词(f_9, f_{10}) 。

6) 假设当前单词 w_0 与前一单词 w_{-1} 在短语句

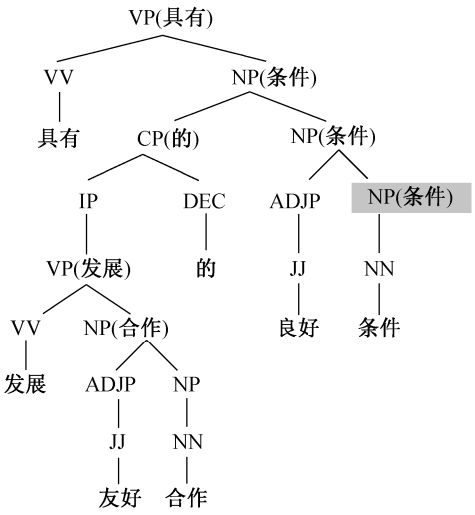

图 2 词汇化句法树的例子

Fig.2 An example of lexicalized syntactic tree

法树中对应的最高同源父节点分别为 N_{w0} 和 N_{w-1}，则 f_{11} 为 N_{w0} 树中的层数是否与 N_{w-1} 相等，例如单词"条件"的 f_{11}=1。

7) N_{w0} 树中的层数与 N_{w-1} 的层数之差是否为1(f_{12})，为 2(f_{13})，大于等于 3(f_{14})，如 w_0="条件"的 f_{12}=否， f_{13}=否，f_{14}=否

8) N_{w0} 树中的层数与 N_{w-1} 的层数之差是否为 −1(f_{15})，为−2(f_{16})，小于等于 3(f_{17})。

对于最大熵模型和条件随机场模型，我们使用[-2,2]的窗口长度对每个单词抽取联合特征，所以每个特征向量都有 5*17 维；而对于 SVM，我们将字符串特征转化为二进制的表示方式，例如

$$f1=\begin{cases}1, \text{word}=W,\\ 0, \text{word}\neq \text{W}。\end{cases}$$

3 基于最大熵的篇章结构和关系的自动学习

当篇章被表示为篇章单元的序列后，我们进一步建立篇章结构树并识别篇章关系。考虑到最大熵模型在分类效果和运行速度上都有良好的表现，我们使用了基于最大熵的参数估计方法。另外引言中提到传统的 PCFG 不适用于篇章结构的推导，因此，本文首先采用了改进后的 CYK 算法进行结构推导(建树)，然后再进行关系识别。

3.1 结构分析的训练和解码

TCT 语料中所标注的复句关系在结构上与图 1 类似，可以形成树状结构，只是并非二叉化的结构。为了方便模型的训练和解码，我们通过添加虚拟节点的方式将其转化为向左二叉化的树，树上的叶子节点为基本篇章单元。在一棵二叉化的篇章结构树上，我们用 $P(1|s_1,s_2)$ 表示跨度 1(s_1)与跨度 2(s_2)存在连接节点的概率，相反 $P(0|s_1,s_2)$ 表示不存在连接节点的概率。所以，整棵树的打分$S(T)$可以通过遍历所有非叶子节点，即遍历所有的中间节点(node)就能得到。$S(T)$可以写成下面的形式:

$$S(T)=\prod_{\text{node}} P(1|\text{n(lchild)},\text{n(rchild)})$$

其中，n(lchild)表示当前 node 的左孩子，n(rchild)表示当前 node 的右孩子，由上式可得，在训练过程中需要估计所有相邻跨度之间是否存在连接节点的概率，相应的在解码过程中，通过对每两个相邻跨度的节点连接的可能性进行打分，我们就可以动态地生成一棵使得$S(T)$取最大值的树 T，从而得到最优化的二叉树。

接下来的问题就是如何估计参数。从图 1 中可知，跨度 2 与跨度 1 连接而未与跨度 3 连接，因此我们可以获得一个有用的信息：跨度 2 在跨度 1 与跨度 3 之间选择了 1，这说明跨度 1、2 具有某种内在联系，使它们更倾向于连接，而跨度 2、3 更倾向于不连接。遍历所有相邻的跨度，抽取连接和非连接的正负样例，然后使用最大熵模型和梯度下降法，就能训练出每两个相邻跨度连接在一起的概率分布。

我们采用 CYK 搜索算法进行解码。传统的 CYK 算法通常是从 PCFG 的推导集合中选取一组得分最高的推导规则，而这里因为我们只关心两个子树是否需要向上生成新子树，因此只对两个相邻子树向上合并进行打分，通过自底向上记录每个新生成子树的最高得分从而取得最佳路径，伪代码如下:

```
for level from 1 to n:    //跨度从 2 到 n
  for col from 0 to n-level: //遍历 level 层所有单元
   Score[level][col]=0;//存储 col~col+level 的最高得分
score=0;
   for k from 0 to level: //遍历所有可能的组合
    l1=level-1-k
    r1=col+1+k
    score+=Score[k][col]+Score[l1][r1];
    score+=Struct(Span(k, col),   Span(l1, r1));
    if score >Score[level][col] :
     Score[level][col]=score;
     Path[level][col]=k;//存储最佳路径，最后从顶向下遍历 Path，构建树结构
```

3.2 关系标注

原始的树结构二叉化以后，关系标签除了 TCT 中的 11 种关系之外，还有构建虚拟节点时所新添的关系，我们用 PX 来表示。对于这个多分类问题，为了与构建树结构的训练解码过程统一起来，我们同样采用最大熵分类器，在找到最佳构建树的路径后立即判断每个节点所控制的两个跨度之间的逻辑语义关系。

3.3 结构和关系学习的特征

3.3.1 子树的支配关系

Soricut 等[3]提出支配(dominance)集合的概念，我们先简单回顾一下支配节点以及支配关系的概念。对于篇章单元 U 在句法树中所对应的子树片

段, 总能找到片段上一个最高的节点 N_H, 使得 U 中的某一个词就是该节点对应的中心词 H, 如图 3 中 U_2 的 H=多, N_H =IP(多)。那么, 除了 N_H 为根节点的篇章单元以外, 其余篇章单元的 N_H 节点的父节点的中心词肯定总是属于另外一个篇章单元, 这个父节点被称为附着节点 N_A, 如 U_2 的 N_A =CP(虽然), 它的中心词属于篇章单元 U_1。这样的现象可以用支配关系 $(U_i,N_H)<(U_j,N_A)$ 来表示, 如(U_2, IP(多))<(U_1, CP(虽然)), 简单表示为 2<1。

这种支配关系还被 Duverle 等[7]用于基于 RST 理论的英文篇章分析, 证明是有用的特征, 我们认为在中文篇章的修辞结构中也有类似的特点, 都是是逻辑语义关系的反映, 可以用来帮助分析篇章关系。

3.3.2 所有特征的集合

表 1 给出了本节用到的所有特征模板, 部分特征受文献[7]的启发。其中 L 表示左孩子, R 表示右孩子, F 表示整个跨度。其中 f_1 ~ f_{25} 为预测结构的特征模板, f_1 ~ f_{13}、f_{20} ~ f_{26} 被用来预测两跨度之间的关系。

4 实验

4.1 实验设置

我们的实验分两部分, 一是篇章切分实验: 1)章结构和关系分析的实验。两部分的实验语料都来自清华汉语树库实时报道部分, 训练集包括以 NEWS 为前缀的编号从 0002 到 0361 的文件, 测试集包括从 0362 到 0376 的文件。清华树库共标注 11 种逻辑关系(包括并列、连贯、递进、选择、因果、目的、假设、条件、转折、注解和流水), 并标注每种关系所控制的语义片段的起始位置和结束位置。我们根据树库中的标记提取出语义单元以及

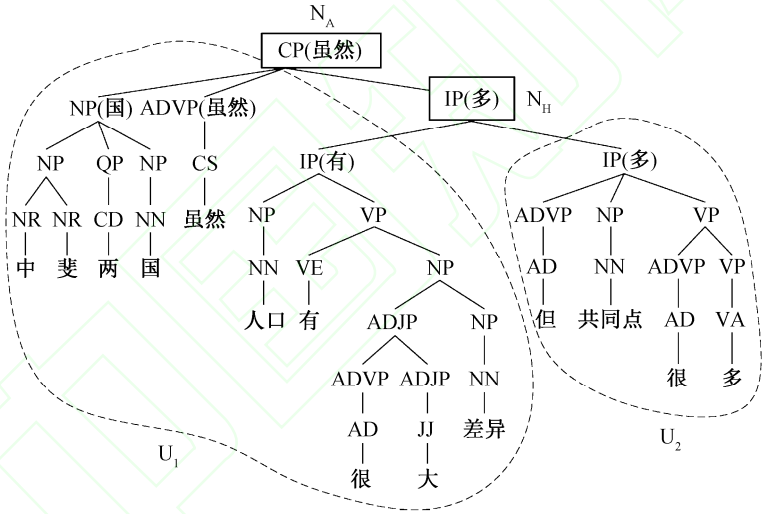

图 3 支配关系示例图

Fig.3 An example of dominance

表 1 篇章结构和关系标注所用到的特征

Table 1 Features for discourse structure and relationship labeling

特征编号	特征含义	特征编号	特征含义
f_1	L 与 R 的支配关系	f_{16}	F 的最高节点 N_H 的中心词
f_2(f_3)	L(R)的附着节点(N_A)(N)的中心词	f_{17}	F 的最高节点 N_H 的句法标记
f_4(f_5)	L(R)的附着节点(N_A)的句法标记	f_{18}	F 的兄弟节点的句法标记
f_6(f_7)	L(R)的最高节点 N_H 的中心词	f_{19}	L 的 N_H 到 F 的 N_H 的距离与 R 的 N_H 到 F 的 N_H 的距离之差之差
f_8(f_9)	L(R)的最高节点 N_H 的句法标记		
f_{10}(f_{11})	L(R)中所包含的篇章单元的个数	f_{20}(f_{21})	L(R)的起始单词
f_{12}(f_{13})	L(R)的 N_H 的层数	f_{22}(f_{23})	L(R)的起始二元单词组
f_{14}	F 的附着节点(N_A)的中心词	f_{24}(f_{25})	L(R)的起始三元单词组
f_{15}	F 的附着节点(N_A)的句法标记	f_{26}	F 是上层节点的左孩子还是右孩子

各单元间的逻辑语义关系。

篇章切分的实验中共抽取 145143 个训练实例，19216 个测试实例，我们比较了最大熵(ME)、支持向量机(SVM)和条件随机场(CRF)这 3 个常用于序列标注任务模型的切分效果。

第二部分的实验中，我们抽出 56949 个训练实例，测试部分分为两组，首先输入已经标注好的切分单元，对输出的二叉化结构和关系进行评价，并且将二叉化结构还原为非二叉树结构，并对其结构和关系进行打分；第二步，输入自动篇章切割的单元，并对最终输出的非二叉树的结构和关系进行打分。

对于结构和关系的打分，我们采用标准的 Parseval 打分矩阵[23]，也就是正确分析的跨度占分析得到的所有跨度的百分比作为准确率 P，正确分析的跨度占参考答案所有跨度的比值作为召回率 R，并且综合考虑准确率 P 和召回率 R 的整合打分为 $F=2PR/(P+R)$。

4.2 实验结果

4.2.1 篇章切分结果

表 2 比较了不同窗口下各种分类器的切分效果，其中[–2, 2]表示选取从位置–2 到位置 2 的窗口。

实验结果表明，使用 SVM，在本测试集上能达到 89.1%的 F 值，比其他两个模型的效果好，但是由于特征空间比较复杂，SVM 往往在训练的时候花费的时间更多。当仅用当前位置的特征时，各个模型的准确率相比于最高值降低了 9.8%～23.4%，召回率相比于最高值降低了 21.0%～29.2%，而效果最好的特征则出现在[–2, 2]以及[–2, 0]的窗口长度上。

我们统计了逗号在测试集标准答案中作为边界出现的频率，大约占 84%，而用 SVM 窗口为[-2, 0]的自动切分结果中，逗号作为边界的比重大约为 88%，与标准答案的比例相差不大。

4.2.2 结构和关系分析结果

输入的篇章单元为人工切分的标准篇章单元时，在测试集上的打分如表 3 所示，其中 Bin-struct 表示二叉化的结构，Bin-struct+label 表示二叉化结构加上关系以后的打分，Normal-struct 表示还原为非二叉化后的树结构，Normal-struct+label 表示还原树结构加上关系后的得分，StdStruct+label 表示当树结构为人工标注的标准结构时的篇章关系得分。

输入人工的篇章切分单元，空白二叉结构能得到 77.4%的 F 值，还原为非二叉结构能得到 79.1%的 F 值，而加入篇章关系后的 F 值均有所下降。若对输入的空白结构为人工标注的正确的结构，能得

表 2 不同分类器不同特征窗口的篇章切分结果

Table 2 Segmentation performance of different classifers and different window lengths

窗口位置	ME			SVM			CRF		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
[–2, 2]	0.874	0.852	**0.863**	0.895	0.887	**0.891**	0.889	0.874	0.881
[–2, 0]	0.868	0.851	0.859	0.900	0.886	**0.891**	0.893	0.879	**0.886**
[–1, 1]	0.872	0.842	0.857	0.900	0.879	0.890	0.885	0.874	0.880
[–1, 0]	0.864	0.839	0.851	0.902	0.879	0.890	0.890	0.879	0.885
0	0.695	0.567	0.625	0.568	0.595	0.702	0.795	0.669	0.727

表 3 输入标准切分单元和自动切分单元的的树结构和关系分析结果

Table 3 Structure labeling and relationship labeling with standard segmented and automatically segmented elementary discourse unit

测试项目	人工切分			自动切分		
	P	*R*	*F*	*P*	*R*	*F*
Bin-struct	0.773	0.774	0.774	0.627	0.622	0.624
Bin-struct+label	0.657	0.658	0.657	0.508	0.504	0.506
Normal-struct	0.835	0.751	**0.791**	0.681	0.640	**0.660**
Normal-struct+label	0.663	0.496	0.568	0.533	0.416	0.467
StdStruct+label	0.875	0.875	0.875	—	—	—

到 87.5%的关系识别 *F* 值。

4.3 结果分析

由于本文的解码算法是自底向上逐层依次打分, 中间每一层的错误都有可能传递到上一层, 为了更清楚的看到本文方法在哪些层次区间更有效, 我们将人工切分的非二叉化的空白结构以及篇章关系分析结果按篇章关系的层数描绘出来, 如图 4 所示。

图中的横坐标表示关系层数, “所占比例”曲线上的点(*x*, *y*)表示关系层数不高于 *x* 的句子所占的比例为 *y*, “结构 *F* 值”上的点(*x*, *y*)表示关系层数不高于 *x* 的非二叉化结构的 *F* 值得分, “关系 *F* 值”的含义与之类似。从图中我们可以看出 TCT 测试集上有 98%的句子所包含的关系层数低于 10, 而当句子关系只有两层或者三层时, 篇章关系的打分达到最高值, 而关系层数不高于 3 时结构分析效果最好。关系层数不高于 6 时的句子占到 93.3%, 结构得分达到 0.81, 这说明在绝大多数复句级别的篇章分析中, 篇章结构能够得到较好的预测。但由于训练数据中特别复杂的句子较少, 数据稀疏, 所以我们的模型对层数较高的篇章特征把握不足, 高层篇章关系分析能力较弱, 有待进一步提升。

另外, 我们统计了输入标准的篇章结构时各个篇章关系的识别情况, 发现递进、因果、条件关系的正确率比较高, 分别是 0.86、0.8 和 0.75, 而目的、并列、流水关系的识别率较低, 分别只有 0.5、0.59 和 0.62。这是因为诸如因果、递进等关系往往会有一些连词作为提示, 如“因为, …, 所以, …, 不仅, …, 还, …”, 比较容易分析。而目的、流水等关系, 常常需要更深入地理解语义并联系上下文才能判断, 当缺少明确的连接词时, 分析起来十分复杂。因此, 我们需要泛化性更强的特征来加强对后者的分析能力。

尽管篇章关系的自动识别率还有待进一步提高, 但是篇章单元的自动识别的效果还是相对可观的, 可用于辅助其他自然语言处理任务。比如在机器翻译中, 鉴于篇章单元具有语义独立性和完整性, 我们可以用篇章单元来约束翻译结果的语义片段, 使其保证完整,加强翻译结果的可信度和可读性。

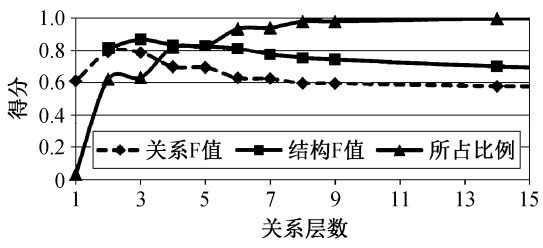

图 4 关系层数与篇章关系标注结果的关系

Fig. 4 Trend of labeling performance with the relationship levels rising

5 结语

本文研究了如何自动分析汉语篇章的篇章结构和篇章关系。对于篇章单元的识别, 我们采用序列标注的方法并且比较了几种常见模型的识别效果; 在篇章树的构建过程中, 我们提出了基于最大熵模型的篇章结构分析方法, 对篇章结构和篇章关系分别建模, 实验结果表明, 当复句中的篇章关系不超过 6 层时, 结构得分能得到 0.81 的 *F* 值, 当复句只有两层或者三层的篇章关系时, 标注关系的 *F* 值能达到 0.79。

在今后的工作中, 我们将进一步筛选有效的特征来减少数据稀疏引起的问题, 另外, 我们将进一步研究 TCT 中其他领域如文学、科技、回忆录等的篇章结构, 并将汉语篇章分析用于自然语言处理的其他任务, 如机器翻译、情感分析等。

参考文献

[1] Mann W C, Thompson S A. Rhetorical structure theory: a framework for the analysis of texts.IPRA Papers in Pragmatics, 1987, 1: 79–105

[2] Mann W C, Thompson S A. Rhetorical structure theory: toward a functional theory of text organization. Text, 1988, 8(3): 243–281

[3] SoricutR,MarcuD. Sentence level discourse parsing using syntactic and lexical information // Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1.Edmonton: Association for Computational Linguistics, 2003: 149–156

[4] Le ThanhH, AbeysingheG, HuyckC. Automated discourse segmentation by syntactic information and cue phrases // Proceedings of the IASTED International Conference on Artificial Intelligence and Applications. Innsbruck, Austria, 2004: 411–415

[5] SubbaR,Di Eugenio B. Automatic discourse segmentation using neural networks // Proceedings of the 11th Workshop on the Semantics and Pragmatics

of Dialogue. Trento, Italy, 2007: 189–190

[6] SubbaR,Di Eugenio B. An effective discourse parser that uses rich linguistic information // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.Boulder: Association for Computational Linguistics, 2009: 566–574

[7] DuVerleD, PrendingerH. A novel discourse parser based on support vector machine classification // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore: Association for Computational Linguistics, 2009:665–673

[8] HernaultH, PrendingerH, duVerleD, etal. HILDA: a discourse parser using support vector machine classification. Dialogue and Discourse, 2010, 1(3):1–33

[9] 乐明. 汉语篇章修辞结构的标注研究.中文信息学报, 2008, 22(4): 19-23

[10] Zhou Yuping, XueNianwen. Pdtb-style discourse annotation of Chinese text // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju: Association for Computational Linguistics, 2012: 69–77

[11] Song Rou, Jiang Yuru, Wang Jingyi. On generalized-topic-based Chinese discourse structure // CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, 2010: 23-33

[12] 周强. 汉语句法树库标注体系. 中文信息学报, 2004, 18(4): 1–8

[13] Prasad R, Dinesh N, Lee A, et al. The penn discourse treebank 2.0 // Proceedings of the 6th International Conference on Language Resources and Evaluation.Marrakech, Morocco, 2008: 2961–2968

[14] Lin Ziheng, Ng H T, Kan M Y. A PDTB-styled end-to-end discourse parser. No. arXiv: 1011.0835. Technical Report TRB8/10, School of Computing, National University of Singapore, August, 2010

[15] Zhou Zhimin, Xu Yu,NiuZhengyu, et al. Predicting discourse connectives for implicit discourse relation recognition // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing: Association for Computational Linguistics, 2010: 1507–1514

[16] 徐凡,朱巧明, 周国栋. 基于树核的隐式篇章关系识别. 软件学报, 2013, 24(5): 1022-1035

[17] Li Yancui, FengWenhe, Zhou Guodong. Elementary discourse unit in Chinese discourse structure analysis // Chinese Lexical Semantics. Berlin: Springer, 2013: 186–198

[18] 姚双云, 胡金柱,舒江波, 等. 篇章连贯语义关系的自动标注方法. Computer Engineering, 2012, 38(7):131-133

[19] Zhou Q. Evaluation report of the third Chinese parsing evaluation: CIPS-SIGHAN-ParsEval-2012 // Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, 2012: 159–167

[20] Li Dongchen, Wu Xihong. Parsing TCT with split conjunction categories // Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, 2012: 174–178

[21] Hugo H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation // Computational Linguistics and Intelligent Text Processing. Berlin: Springer, 2010: 315–326

[22] Jiang Yuru, Song Rou.Topic structure identification of PClause sequence based on generalized topic theory // Natural Language Processing and Chinese Computing. Berlin: Springer, 2012: 85–96

[23] Ezra B, Abney S,Flickinger D, et al. Procedure for quantitatively comparing the syntactic coverage of English grammars // Proceedings of the workshop on Speech and Natural Language. Association for Computational LinguisticsAsilomar. California, 1991: 306–311