文章编号：1003-0077(2019)02-0001-07

基于序列标注的引语识别初探

贾泓昊，罗智勇

（北京语言大学 信息科学学院，北京 100083）

摘 要：句间引用关系自动识别是篇章分析中一项重要内容。句间引用关系影响着对句群篇章的分析，而目前自然语言处理中对引用这一句间关系的研究较少。句间引用关系主要体现在引语中的引用句上。引语由引导句和引用句组成，一般分为直接引语和间接引语，其中间接引语的识别难度最大。引导句和引用句相对位置不定、不同领域语料的引语与非引语比例极不均衡等进一步增加了引语自动识别的难度。该文主要尝试对引用这一句间关系进行初步探索，采用条件随机场(CRF)以及双向长短期记忆网络与条件随机场相结合(BLSTM-CRF)的方法对引语进行自动识别，并引入引导句中管领词特征进行实验对比。实验结果表明，CRF 模型和 BLSTM-CRF 模型对引语的识别精确率分别达到 85.49%和 80.19%，F 值分别达到 78.75%和 79.60%。

关键词：引语的识别；序列标注；条件随机场；双向长短期记忆网络

中图分类号：TP391 **文献标识码**：A

A Study on Quotation Recognition Based on Sequence Labeling

JIA Honghao, LUO Zhiyong

(School of Computer Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: The automatic recognition of inter-sentence quotation relationship is a valid issue in discourse analysis. The quotation relationship between sentences influences the analysis of sentence groups. At present, there are few studies on the relationship between quotations in natural language processing. This paper attempted to make a preliminary exploration of the relationship between quoted sentences and studied the identification of quotation with conditional random fields(CRF) and Bidirectional Long Short-Term Memory network Enhanced CRF (BLSTM-CRF). It introduces the governors in the leading sentence into the model. The experimental results show that CRF model performs better with 85.49% in precision, and BLSTM outperforms with 79.60% in F-value.

Keywords: quotation recognition; sequence labeling; CRF; BLSTM

0 引言

随着自然语言处理技术的飞速发展，句群篇章语义的分析需求日益增大。作为句群篇章语义分析的重要内容，句间关系也受到越来越多的关注，然而在自然语言处理中关于引用这一句间关系的研究较少。句间引用关系的研究主要集中在语言学方面。

句间引用关系主要体现在引语中的引用句上。引语分直接引语和间接引语。从结构上看，直接引语与间接引语均由引导句和引用句两部分组成。在引导句中，起关键作用的言说动词对引用句有一定的管领作用，因此徐赳赳[1]将引导句的言说动词叫管领词。引用句是被引用的部分，表明句间关系。宋柔[2]也在小句复合体的理论中将引语中的引用句叫做封闭语段，例如：

例 1 不过要是有人问我："你最喜欢什么动物？"(小说《邢老汉和狗的故事》)

例 2 我认为，这绝不是画家在故作玄虚，也不是虚构的人格化的动物形象，一定是画家对实有其狗的小友的纪念。(小说《邢老汉和狗的故事》)

例 1 是直接引语，由引导句"不过要是有人问我"和引用句"你最喜欢什么动物"组成，管领词为"问"。宋柔在小句复合体理论中认为句子是有层级

收稿日期：2018-09-29 **定稿日期**：2018-10-29

基金项目：北京市哲学社会科学规划研究基地项目(13JDZHB005)

结构的,指出引用句“你最喜欢什么动物?”相对于引导句“不过要是有人问我”结构上是封闭的,称为封闭语段。例2是间接引语,包括引导句“我认为”和引用句“这绝不是画家在故作玄虚,也不是虚构的人格化的动物形象,一定是画家对实有其狗的小友的纪念”,管领词为“认为”,“这绝不是画家在故作玄虚,也不是虚构的人格化的动物形象,一定是画家对实有其狗的小友的纪念。”是封闭语段。

本文主要识别句间引用这一关系,也就是识别表明引用关系的引用句,如例1,识别出“你最喜欢什么动物”是引用句,也就识别出例1是引语。按照封闭语段的定义,引用句相当于外部是封闭的,可以当作一个总体来识别。因此,在本文中,我们主要识别的是引语中表明引用关系的引用句,引用句的识别主要有以下3个问题。

(1) 间接引语难以界定。

如上面例2:我认为,这绝不是画家在故作玄虚,也不是虚构的人格化的动物形象,一定是画家对实有其狗的小友的纪念。(小说《邢老汉和狗的故事》)。这种间接引语没什么特别区分标志,人在学习时需要学习相关语法知识,区分其与一般陈述句的区别。

(2) 引导句和引用句相对位置不定。例如:

例3 “我现在向你补求,行不行?”好像一切没恋爱过的男人,方鸿渐把“爱”字看得太尊贵和严重,不肯随便应用在女人身上。(小说《围城》)

例4 辛楣也笑道:“孙小姐这房间住得么?李梅亭更住不得……”正说着,听得李顾那面嚷起来。(小说《围城》)

例5 孙小姐凑上去瞧,不肯定地说:“这像是西药。”(小说《围城》)

如例3所示,引用句“我现在向你补求,行不行?”在句子的开头;例4的引用句“孙小姐这房间住得么?李梅亭更住不得……”在句子的中间;例5的引用句“这像是西药。”在句子的结尾。引用句没有固定的位置,无疑增加识别的难度。

(3) 引语分布不均。

一方面,不同领域引语比例相差较大(详细介绍见表3);整体而言,引语占比较少。我们对实验语料(详细介绍见表3)进行分析,总计13 370句语料,含有引语的句数是2 766,占比20.69%。这是一个不平衡的分类或标注问题。

问题(1)是引语的本身问题。问题(2)需要我们采用的方法对位置信息不敏感。问题(3)需要我们在模型计算的时候采用负采样的方法,使数据尽量平衡。

本文尝试采用序列标注的方法,提出基于条件随机场(CRF)以及深度神经网络与条件随机场相结合(BLSTM-CRF)两种模型对引语进行识别,同时引入管领词特征来提高识别的效率。

本文第1节介绍句间关系与序列标注的相关研究;第2节介绍条件随机场模型;第3节介绍条件随机场与双向长短期记忆网络相结合模型;第4节介绍实验及结果;第5节总结并提出下一步工作。

1 相关研究

目前在自然语言处理领域针对引语自动识别的研究较少。典型的篇章句间关系语料有以下两种:基于RST理论[3]的修辞结构理论树库(rhetorical structure theory discourse treebank)[4]和基于PDTB体系的宾州篇章树库(Penn discourse tree bank)[5],它们采用不同的关系类型体系和标注标准[6]。但是这些均是英文方面的研究,张牧宇[7]采用自身提出的中文句间关系理论,对中文句间关系进行研究,但这些都不是特别针对引语识别的研究。

本文首次尝试对引用这一句间关系进行研究分析,找出引用关系中的引用句,采用序列标注这一方法对引用句进行自动识别。目前,序列标注模型主要包括:隐马尔可夫模型(HMM)、最大熵马尔可夫模型(MEMM)[8]和条件随机场(CRF)[9]。近年来,随着深度学习技术的发展,研究者将深度神经网络引入到序列标记任务中,具有代表性的研究工作包括:Collobert[10]使用卷积神经网络模型来解决序列标注中通用命名实体识别的问题;Huang[11]首次采用将双向长短期记忆网络与条件随机场结合的方法来进行序列标注,并在词性标注、命名实体识别等任务上取得良好效果。

本文尝试采取条件随机场(CRF)、双向长短期记忆网络与条件随机场相结合(BLSTM-CRF)这两种方法对引语进行识别,同时引入管领词特征来提高识别的效率。实验结果表明,CRF模型和BLSTM-CRF模型对引语的识别精确率分别达到85.49%和80.19%,F值分别达到78.75%和79.60%。

2 条件随机场(CRF)

2.1 条件随机场介绍

条件随机场(CRF)是给定一组输入随机变量条

件，求另外一组输出随机变量的条件概率分布模型；其特点是假设输出随机变量构成马尔科夫随机场。本文用到的是线性条件随机场，直接对输入数据进行处理，获得全局最优的标记序列。

2.2　引语的识别建模

如前所述，引语可以作为一个整体进行识别，本文将引语识别问题转化为序列标注问题。该序列标注模型可以定义为：给定一个长度为 n 的句子 $X=\{x_1,x_2,\cdots,x_n\}$，从所有可能的标记序列中挑出最可能的标记序列 $Y=\{y_1,y_2,\cdots,y_n\}$，最终从获得的标记序列中还原引语的位置。

一个词语在句子中的标记方式有四种：B 代表一个引用句的开头，I 代表引用句中间的词语，E 是一个引用句中的最后一个词语，O 则是其他非引用句的词语。一个简单的面向引语识别的序列标记实例如表 1 所示。

表 1　引语识别实例

句子：	她	说	今天	很	开心	。
标注结果	O	O	B	I	E	O
识别结果	今天很开心					

2.3　特征模板

在神经网络中用到 CRF 是不需要采用特征模板的，因为神经网络会自己学习里面的规律，例如，下面介绍的 BLSTM-CRF 就是通过两层 BLSTM 来学习句子的内部规律。但是直接用条件随机场的方式来进行序列标注需要特征模板。条件随机场通常采用文本窗口的方式定义特征，特征定义方式以某字符相对于当前位置在文本中的偏移位置来表征。本文使用了 12 种字符特征模板，包括一元字符特征模板 C_0（当前字符）、C_{-1}（当前字符向前第一个字符）、C_1（当前字符向后第一个字符）、C_{-2}（当前字符向前第二个字符）、C_2（当前字符向后第二个字符），二元字符特征模板 $C_{-2}C_{-1}$、$C_{-1}C_0$、C_0C_1、C_1C_2 和三元字符模板 $C_{-2}C_{-1}C_0$、$C_{-1}C_0C_1$、$C_0C_1C_2$。

在本模型中，我们使用两套模板，一套考虑管领词特征，另一套未考虑。未考虑管领词的引语识别标记序列实例如表 1 所示。考虑管领词特征时，我们在语料中多加入一列特征，管领词部分标记为 1，非管领词部分标记为 0。引入管领词的引语识别标记序列实例如表 2 所示。

表 2　引入管领词特征引语识别实例

句子：	她	说	今天	很	开心
管领词	0	1	0	0	0
标注结果	O	O	B	I	E
识别结果	今天很开心				

3　双向长短期记忆网络结合条件随机场（BLSTM-CRF）模型

3.1　模型介绍

图 1 为本文的 BLSTM-CRF 模型框架。模型框架主要由以下几个网络层组成：第一层为词向量表示层（word embedding）；第二层为循环神经网络层，包含两层双向 LSTM 循环神经单元（BLSTM）；最后一层为 CRF 层。

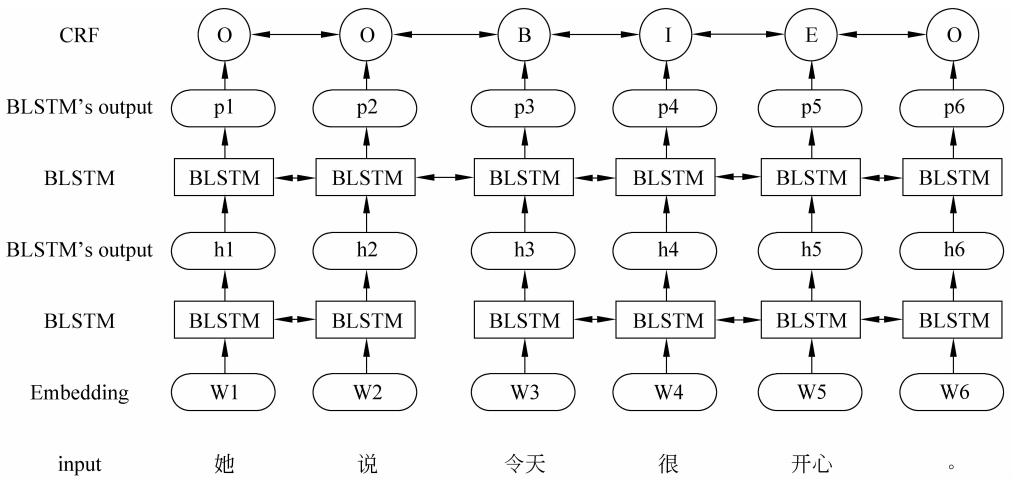

图 1　引语识别的 BLSTM-CRF 模型

首先，我们通过查询词向量表将输入的语句转换为相应的词向量序列；同时，除词向量外，我们还引入其他特征向量，如管领词向量，将这些特征向量和词向量拼接，作为模型的输入；然后，将上述词语特征向量序列输入循环神经网络层；最后，模型将循环神经网络层在每一个时刻的输出，作为 CRF 的输入序列，生成最优的标记序列。

3.2 词向量层

本文的词向量有两种处理方式。一种是直接查询词向量，将输入的语句转化为相应的词向量。但是在对引语进行自动识别探索时发现，管领词对于引用识别边界具有很大的提示作用。因此，本文从大规模文本中收集和整理出来引导语中常用的高频管领词表，在进行词向量处理的时候引入相应的管领词向量。如图 2 所示。

对于管领词向量我们设计如下处理方法：将原有词向量维度增加一个维度，非管领词部分赋值为 0，管领词全部赋值为 1，并且在训练过程中不会改变其值，标记为不可训练。

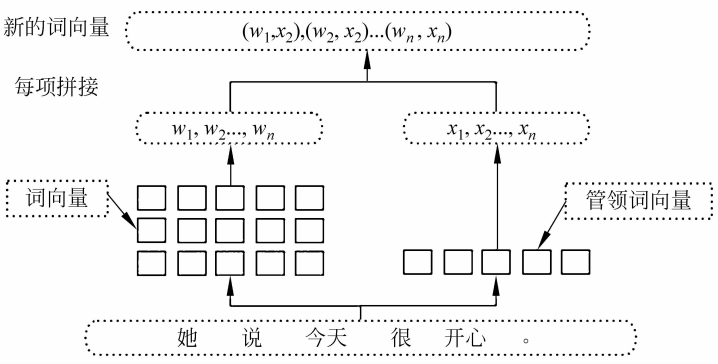

图 2 引入管领词的词向量层

3.3 双向长短期记忆网络(BLSTM)层

LSTM(long short-term memory)是长短期记忆网络，是一种特殊的 RNN 类型，可以学习长期依赖信息。LSTM 模型通过特殊设计门结构(遗忘门、输入门及输出门)来控制长期状态。

引语的识别需要充分用到上下文的信息，我们采用双向 LSTM(即 BLSTM)。双向 LSTM 对句子能够从左到右进行顺序计算以及从右到左进行逆序计算，得到两种不同的隐层表示，然后合并成最终的隐层输出。

3.4 条件随机场(CRF)层

此模型将两层 BLSTM 处理后的隐层输出结果当做 CRF 的输入，获得最优标记序列。

输入一个长度为 l 的句子 $S=\{w_1,w_2,\cdots,w_l\}$，定义两层 BLSTM 的输出概率矩阵 $\boldsymbol{P}_{l*k}$，其中 k 是输出标签的个数，在我们的问题中 $k=4$。$P_{i,j}$ 是指第 i 个词语被标记为第 j 个标签的概率。对于一个待预测的标签序列：$y=\{y_1,y_2,\cdots,y_l\}$，有如下定义，如式(1)所示。

$$f(x,y)=\sum_{i=0}^{l}(A_{y_i,y_{i+1}}+P_{i,y_i}) \tag{1}$$

其中，$\boldsymbol{A}$ 是状态转移矩阵，$A_{y_i,y_{i+1}}$ 表示从标签 y_i 转移到标签 y_{i+1} 的概率。通过求得最大的 $f(s,y)$，即可得到最佳的输出标签序列。这里引入的 CRF，其实只是对输出标签二元组进行建模，然后使用动态规划进行计算即可，最终根据得到的最优路径进行标注。

4 实验

4.1 评测数据集

本文测试数据集来自北京语言大学中文小句复合体标注语料，主要分为 3 个领域：百科、小说、新闻。语料的详细信息如表 3 所示。

表 3 数据集介绍

语料类别	总句数	含引语句数	引语比例/%
百科	3 322	152	0.30
新闻	1 007	104	10.32
小说	9 041	2 510	27.76
总计	13 370	2 766	20.69

以上语料集中每一个小句都是按照小句复合体的理论来划分的。在语料处理方面，为能够清楚地表示语料中待识别的引语，我们采用 BIOE 标记的

方式来标记引语。我们的标记规则如下：B 代表一个引用句的开头，I 代表引用句中间的词语，E 是一个引用句中的最后一个词语，O 则是其他非引用句的词语。将每个语料中的含引语句子、不含引语句子均按照 8∶2 的比例划分训练集、测试集。

4.2　实验结果及分析

4.2.1　百科语料

百科语料引语识别的实验结果如表 4 所示。百科语料中引语中 90%为间接引语。可以看出，在不引入管领词的情况下，CRF 在各项指标上的结果均高于 BLSTM-CRF，说明 BLSTM-CRF 对引语内部规律的学习能力较弱，不如直接给特征模板学习的效果显著。管领词的引入，对于模型 CRF，在引语的识别方面有显著的提升，精确率由 37.50%提升到 62.50%，但是在模型 BLSTM-CRF 上精确率只有 0.9%的提升。百科语料中，在 CRF 中直接引入管领词特征比在 BLSTM-CRF 中引入管领词向量效果更加明显。

表 4　百科语料引语识别结果

模型	引语		
	精确率/%	召回率/%	*F* 值/%
CRF	37.50	7.32	12.24
CRF＋管领词	62.50	29.42	40.00
BLSTM-CRF	6.43	7.24	6.81
BLSTM-CRF＋管领词	7.37	10.53	8.67

4.2.2　新闻语料

新闻语料引语识别的实验结果如表 5 所示。新闻语料中，引语全为间接引语。在不引入管领词的情况下，CRF 对于新闻语料中引语识别的精确率很高，同时引入管领词对于间接引语识别的提升效果很明显，召回率由 9.09%提升到 31.82%，*F* 值由 16.67%提升到 48.28%。BLSTM-CRF 对于间接引语的识别精确率很低，只有 22.86%，引入管领词后，召回率提升了 17.31%，*F* 值提升了 9.71%。新闻语料中，引入管领词，对 CRF 与 BLSTM-CRF 均有一定程度的提升。

表 5　新闻语料引用识别结果

模型	引语		
	精确率/%	召回率/%	*F* 值/%
CRF	100	9.09	16.67
CRF＋管领词	100	31.82	48.28
BLSTM-CRF	22.86	15.38	18.39
BLSTM-CRF＋管领词	24.64	32.69	28.10

4.2.3　小说语料

小说语料引语识别的实验结果如表 6 所示。小说语料中引语中大多数均为直接引语，占比 90.00%左右。在不引入管领词的情况下，CRF 模型比 BLSTM-CRF 模型在各项指标上表现效果更好，说明 BLSTM-CRF 对于引语的区分能力较弱，对于区分规律的学习不如 CRF 给定的特征模板。引入管领词后，CRF 模型在精确率、召回率、*F* 值上有 1%左右的浮动，而 BLSTM-CRF 模型在精确率、*F* 值上分别有 12.94%、7.58%的提升，在召回率上只有 0.51%的提升。小说语料中，引入管领词，对 BLSTM-CRF 有一定程度的提升。

表 6　小说语料引语识别结果

模型	引语		
	精确率/%	召回率/%	*F* 值/%
CRF	85.17	77.13	80.95
CRF＋管领词	86.18	76.95	81.30
BLSTM-CRF	60.49	76.02	67.37
BLSTM-CRF＋管领词	73.43	76.53	74.95

4.2.4　全部语料

全部语料引语的识别实验结果如表 7 所示。在全部语料中，在不引入管领词的情况下，CRF 模型比 BLSTM-CRF 模型在各项指标上表现效果更好。说明 BLSTM-CRF 对于引语的区分能力较弱，对于区分规律的学习不如 CRF。在 CRF 中，引入管领词后召回率、*F* 值分别仅有 1.64%、0.52%的提高，提升效果不是很明显。而在 BLSTM-CRF 中，在精确率、召回率、*F* 值上均有 10%左右的提升，说明引入管领词向量对于 BLSTM-CRF 有很大程度的提升。

表 7　全部语料结果

模型	引语		
	精确率/%	召回率/%	*F* 值/%
CRF	85.49	72.10	78.23
CRF＋管领词	84.49	73.74	78.75
BLSTM-CRF	70.28	67.97	69.10
BLSTM-CRF＋管领词	80.19	79.00	79.60

4.2.5 实验分析

为进一步了解在 CRF 模型与 BLSTM-CRF 模型中，管领词对引语标注结果的影响，考虑到样本大小，本文对全部语料进行分析，在 BLSTM-CRF 中，得到 CRF 层的输入，对管领词与 B 标签词语进行相似度计算，我们发现有以下规律：

① 在两个神经网络模型中均被标注正确的引语中，引入管领词向量后，管领词与 B 标签词的相似度降低；

② 引语在原 BLSTM-CRF 模型中标注错误，引入管领词向量后，标注正确，管领词与 B 标签词相似度降低。

对于 BLSTM-CRF 模型而言，引入管领词特征能够对学习内部规律起到一定的指导作用，降低管领词与 B 标签词的相似度，可以避免它们打上同一标签或者关联标签，从而提升模型效果。

在 CRF 中，我们取出一个考虑了管领词特征的具体特征模板%x[-2,1]/%x[-1,1]/%x[0,1]（接下来皆叫特征模板 C）来进行分析，计算在 B 标签词取不同标记时的权重变化情况，正确标注例句如下：

例 6 他拍拍身边的椅子，说，谢兰英，你靠着我坐。（小说《邢老汉和狗的故事》）

计算“谢兰英”在不同标记时，特征模板 C 权重随迭代次数变化值如图 3 所示。

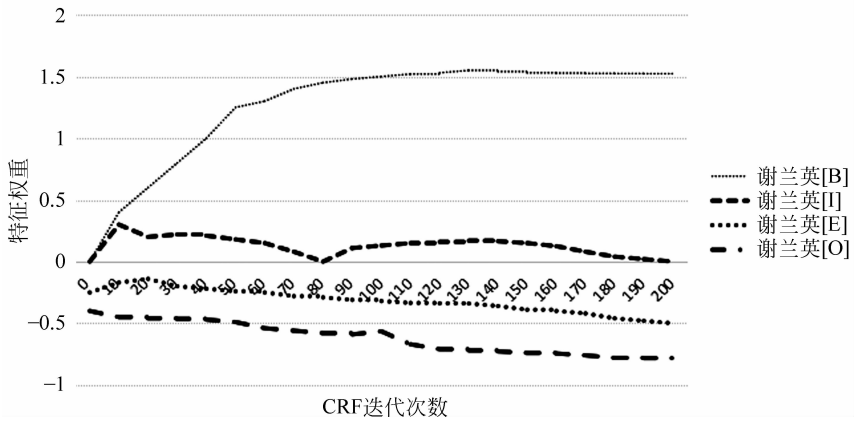

图 3 引入管领词特征“谢兰英”特征权重变化图

在 CRF 中，其他标记正确 B 标签词特征权重变化趋势基本与图 3 相符合。从图中可以看出，随着迭代次数增加，特征 C 使“谢兰英”标记为 B 的权重越来越大，说明特征 C 影响着正确的标记结果。而不引入管领词特征时，对应的特征模板%x[-2,0]/%x[-1,0]/%x[0,0]（接下来皆叫特征模板 C1），也能指导学习“说”与“谢兰英”之间的关系，此时，特征模板 C1 权重随迭代次数变化值如图 4 所示。

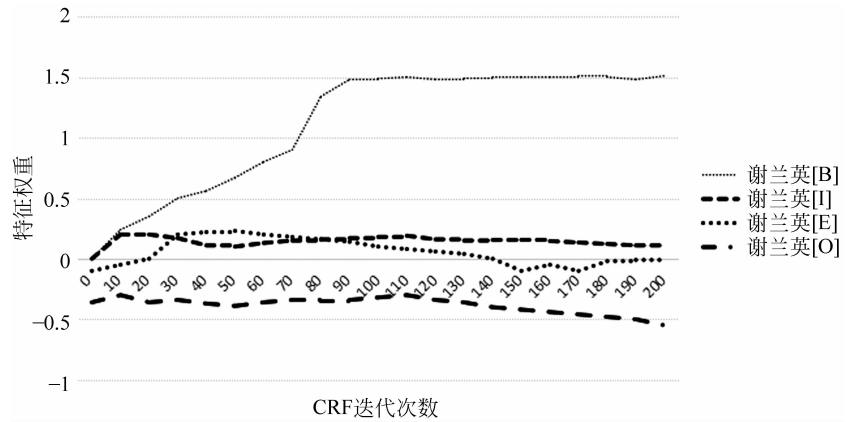

图 4 不引入管领词特征“谢兰英”特征权重变化图

通过图 3 可知，CRF 的特征模板可以指定学习上下文词语之间的关系，管领词“说”对于引用开始位置“谢兰英”有指示作用，而特征模板可以指定学习“说”与“谢兰英”的特征关系，并且特征权重在学习的过程中会逐渐变大，学习到其中的内部规律。通过图 3、图 4 可知，对于 CRF 来说，不管是否引入

管领词特征，都可以通过特征模板直接或间接地学习管领词的管领作用，所以总体上效果提升不大。

5　结论及下一步工作

引语识别是进行小句复合体乃至篇章分析的重要环节。本文提出了基于序列标注的引语识别任务，并通过 CRF、BLSTM-CRF 两种方法分别对引语进行了识别实验，取得了初步效果。同时测试结果表明，引入管领词特征对引语识别具有重要作用。

目前，本文提出的方法没有区分引语中嵌套的引语，例如：

例 7　他说：孔子说，三人行，必有我师。（自拟）

其中“三人行，必有我师”是引用句“孔子说，三人行，必有我师。”中嵌套的引用句。如何从引用句中递归地识别引用句是下一步的研究工作。另外，如何进一步提高 CRF、BLSTM-CRF 对间接引语的识别性能，也是今后的研究内容之一。

参考文献

[1] 徐赳赳. 叙述文中的直接引语分析[J]，语言教学与研究，1996(1)：52-66.

[2] 宋柔. 小句复合体的理论研究和应用. [DB/OL]. http://2011. gdufs. edu. cn/info/1070/2085. htm，2017-11-13.

[3] William Mann，Sandra Thompson. Rhetorical structure theory：Toward a functional theory of text organization[J]. Text，1988，8(3)：243-281.

[4] Carlson L，Marcu D，Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[M]. Springer Netherlands，2003：85-112.

[5] R Prasad，et al. The Penn discourse Treebank 2. 0 [C]//Proceedings of LREC 2008，2008.

[6] A AlSaif，K Markert. The leeds arabic discourse Treebank：Annotating discourse connectives for arabic [C]//Proceedings of LREC，2010：2046-2053.

[7] 张牧宇，等. 中文篇章级句间语义关系识别[J]. 中文信息学报，2013，27(6)：51-57.

[8] A McCallum，D Freitag，F Pereira. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of ICML，2000：591-598.

[9] J Lafferty，A McCallum，F Pereira. Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of ICML，2001：282-289.

[10] Collobert R，et al. Natural language processing (almost) from scratch. [J]. Journal of Machine Learning Research，2011，12(1)：2493-2537.

[11] Huang Z，Xu W，Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv：1508. 01991，2015.

[12] 李航. 统计机器学习法[M]. 北京：清华大学出版社，2012：191-209.

[13] Xuezhe Ma，Eduard Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNsCRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics，2016：1064-1074.

[14] Libin Shen，Giorgio Satta，Aravind Joshi. Guided learning for bidirectional sequence classification[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics，2007：760-767.

[15] Sun X. Structure regularization for structured prediction：theories and experiments[J]. Advances in Neural Information Processing Systems，2014，3：2402-2410.

[16] Kaisheng Yao，et al. Spoken language understanding using long shortterm memory neural networks[C]//Proceedings of the IEEE SLT，2014，12：189-194.

[17] 李丽双，郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报，2018，32(1)：116-122.

[18] Yao L，et al. Biomedical named entity recognition based on deep neutral network[J]. International Journal of Hybrid Information Technology，2015，8(8)：279-288.

贾泓昊(1994－)，硕士，主要研究领域为自然语言处理。

E-mail：honghaojia@foxmail. com

罗智勇(1975－)，通信作者，副教授，硕士研究生导师，主要研究领域为语言信息处理、机器学习。

E-mail：luo_zy@blcu. edu. cn