计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140620
52(9):2114-2122,2015

基于马尔可夫逻辑的中文零指代消解

宋 洋[1] 王厚峰[1,2]
[1](北京大学计算语言学研究所 北京 100871)
[2](计算语言学教育部重点实验室(北京大学) 北京 100871)
(magicsongyang@163.com)

Chinese Zero Anaphora Resolution with Markov Logic

Song Yang[1] and Wang Houfeng[1,2]
[1](*Institute of Computational Linguistics*, *Peking University*, *Beijing* 100871)
[2](*Key Laboratory of Computational Linguistics*(*Peking University*), *Ministry of Education*, *Beijing* 100871)

Abstract Chinese zero anaphora resolution includes two subtasks: zero pronoun detection and zero anaphora resolution, which are correlated with each other. Zero pronoun detection means to recognize all the zero anaphors in a given text, which mainly include null subject or null object, and exist widely in Chinese, Japanese and Italian. Zero anaphora resolution means to determine the antecedent for each recognized zero anaphor, which has already appeared as a noun, pronoun or common noun phrase before the detected zero anaphora in the previous text. Traditional methods to solve Chinese zero anaphora resolution problem generally employ some common-used learning features to construct independent classifiers for zero pronoun detection and zero anaphora resolution, but it cannot capture association relationship between these two subtasks, e.g. recognized zero anaphora must be resolved or the one to be resolved must be zero anaphora and so on. In our method, these two subtasks are combined into a unified machine learning framework with Markov logic to make joint inference and joint learning. We use local formulas to describe zero pronoun detection and zero anaphora resolution respectively, and use global formulas to represent the association relationship between these two subtasks. We find that joint learning model which makes learning with inference can acquire more effective feature weights than independent learning model which just makes learning without inference. Experimental results on OntoNotes3.0 Chinese dataset show that our joint learning model can achieve better results compared with independent learning model and other baseline methods.

Key words Markov logic networks; Chinese zero anaphora resolution; zero pronoun detection; joint learning; global rule; local rule

摘 要 中文零指代消解问题包括零指代项的识别和零指代项的消解2个相互关联的子任务.传统的方法在解决该问题时,往往不考虑2个子任务间的关联关系,比如识别出的零指代项必须被消解以及发生消解的必须是零指代项等约束.基于马尔可夫逻辑网络模型可以将零指代项的识别和零指代项的消解2个子任务融合在统一的机器学习框架下进行联合推断与联合学习,采用局部规则分别针对零指代

收稿日期:2014-07-09;**修回日期**:2014-11-20
基金项目:国家"八六三"高技术研究发展计划基金项目(2015AA015402);国家自然科学基金项目(61370117,61333018);国家社会科学基金重大项目(12&ZD227)
通信作者:王厚峰(wanghf@pku.edu.cn)

项的识别和消解进行预测,采用全局规则描述这 2 个子任务间的关联关系. 基于 OntoNotes3.0 的中文数据集上的实验结果显示,基于马尔可夫逻辑网络的联合学习模型相比于独立学习模型以及多个 baseline 方法能够获得更好的实验效果.

关键词　马尔可夫逻辑网络;中文零指代消解;零指代项识别;联合学习;全局规则;局部规则

中图法分类号　TP301

指代是表示语篇衔接关系的一种重要手段. 指代消解则是理解文本、构建上下文关系、特别是建立实体之间的共指关系的基础;在信息抽取中,指代消解是一个重要的先期性工作. 零指代是一种特殊形式的指代,广泛存在于中文、日文以及意大利文等多种语言中. 与一般指代消解不同,零指代消解需要首先识别零指代项,然后再消解指代项. 本文针对中文零指代消解问题展开研究.

目前的中文零指代消解大多是将零指代项的识别与零指代项的消解这 2 个子问题分开来进行的,如 Zhao 和 Ng[1]的工作以及孔芳等人[2]的工作. 这 2 个子问题是可以关联起来统一建模的,也有相关的研究报道. 例如,Iida 等人[3]基于整数线性规划的思想将零指代项的识别与消解组合在同一个框架下进行联合推断;而采用整数线性规划模型的目的在于,该模型便于实现对 2 个子问题的 2 类预测变量构建相互的约束关系,从而能够在一定约束限制条件下进行目标函数的极值求解. 实验结果也显示,该方法取得了不错的效果. 然而,这一模型虽然也可以称为是"联合"的,但其"联合"仅仅是推断预测变量时的联合,即在一定的约束限制条件下对多类有相互关联的预测变量同时进行推断. 因此,这类模型的准确称法应该是"联合推断"模型,而多个子问题的学习却是"独立"进行的.

本文在马尔可夫逻辑网络的框架下研究中文零指代消解问题,目的是能够基于一个统一的机器学习框架,将零指代项的识别与零指代项的消解 2 个子问题不只进行"联合推断",也同时进行"联合学习",采用一个真正意义上的"联合模型"来解决该问题. 由于马尔可夫逻辑网络框架在特征描述方面有着良好的表达能力,可以将特征与约束采用一阶谓词逻辑规则的形式进行刻画. 局部规则用于刻画不同子问题预测变量的决策,全局规则用于刻画不同子问题预测变量之间的约束限制关系,从而可以将中文零指代问题以不同谓词间的逻辑关系进行清晰地表达.

1　中文零指代消解概述

1.1　问题介绍

零指代[1,4]作为一类特殊的指代现象,广泛存在于自然语言的文本中,特别是在汉语文本中. 下面给出了 2 个例句(例句中所标注的不同表述,上标代表其所指向的实体标号,下标代表其在所处文本段中的顺序标号):

例 1. [吉林省]$^1_1$ 主管经贸工作的副省长全哲洙说:"[* pro *]$^1_2$ 欢迎国际社会同[我们]$^1_3$ 一道,共同推进图们江开发事业,促进区域经济发展,造福东北亚人民."

例 2. 去年初浦东新区诞生的中国第一家[医疗机构药品采购服务中心]$^2_1$,正因为[* pro *]$^2_2$ 一开始就比较规范,[* pro *]$^2_3$ 运转至今,成交药品一亿多元,没有发现一例回扣.

在上述例子中,[* pro *]代表零元素,也就是缺省成分. 该标记在中文宾州树库(Chinese Treebank)中有所定义,用于标识句子中缺省的主语或宾语. 关于不同缺省成分所采用的不同标记类型将在 3.1 节中详细说明. 在例 1 中,双引号后面的缺省部分[* pro *]$^1_2$ 指代了句首出现的[吉林省]$^1_1$ 以及后文出现的[我们]$^1_3$. 例 2 节选自一篇新闻文稿,缺省部分[* pro *]$^2_2$ 以及[* pro *]$^2_3$ 均指代前文出现的名词短语[医疗机构药品采购服务中心]$^2_1$. 在这 2 个例子中的零元素都出现在主语位置,实际的统计表明,主语的缺省且出现在谓语动词之前的零元素占绝大多数.

关于零指代的定义有很多. 本文引用陈平[5]于 1987 年发表在《中国语文》上的定义,即"如果从意思上讲句子中有 1 个与上文中出现的某个事物指称相同的所指对象,但是从语法格局上看该所指对象没有实在的词语表现形式,我们便认定此处用了零指代".

中文零指代消解可以分为 2 个子问题:1)零指代项的识别;2)零指代项的消解. 本文着重研究中文的零指代问题,探讨将 2 个子问题联合起来在一个

统一的框架下求解.从计算模型的角度来看,其方法同样也适用于其他语言中的零指代消解.

1.2 研究现状

很多相关零指代的研究是从认知的角度[6]或语言学的角度[7]来考虑的,本文研究零指代的消解问题,仅从计算的角度介绍零指代问题的研究现状和最新进展.

台湾的 Yeh 和 Chen[4]基于中心理论[8]进行了中文零指代的相关研究,他们手工构建了消解规则,在人工标注的数据集上对零代词的识别与消解都进行了实验,取得了不错的效果.

Converse[9]基于中文宾州树库构建了中文零指代语料库,并根据 Hobbs 算法[10]采用一个规则系统来解决零指代的消解问题,该作者所构建的中文零指代语料库被广泛使用,其后大多数零指代消解的工作都是基于中文宾州树库进行的.不过,该文对于零指代项的识别与检测问题并没有更多地进行讨论.

Zhao 和 Ng[1]首次采用机器学习方法研究了中文零指代消解问题.他们基于 CTB 3.0 数据,选取了其中 205 篇文档,重新构建了中文零指代语料库,针对零指代项的识别以及零指代项的消解,分别构建了词法、语法以及句法特征,并基于 Weka(http://www.cs.waikato.ac.nz/ml/weka/)提供的 J48 决策树分类器进行学习与预测.实验结果显示,他们的方法超过了当时已有的基于启发式规则的方法.

苏州大学的孔芳和周国栋[2]于 2010 年发表在 EMNLP 上的工作尝试采用树核函数的方法以流水线的方式逐次解决零指代消解中的 3 个子任务:零元素检测、零指代项识别以及候选先行语的判断.实验结果显示基于树核函数的方法优于基于平面特征的方法.同时,他们采用的语料库来自于更新版本的 CTB 6.0 数据集,从中选取了 100 篇文档进行训练和测试,并通过上下文敏感的卷积树核(context-sensitive convolution tree kernel)函数来计算句法树中 2 个零指代项的相似度,采用的工具包是 SVM Light (http://svmlight.joachims.org).

除了中文之外,意大利文和日文等语言也同样存在零指代问题.Iida 等人[11]和 Taira 等人[12]的工作涉及到了相关语言,特别是,Iida 等人[3]于 2011 年发表在 ACL 的论文中,采用整数线性规划的解码方法,将零指代项的识别与零指代项的消解 2 个子任务整合在一个框架下进行联合推断并同时应用于意大利文和日文的数据集中,均取得了不错的效果.

2 中文零指代消解模型

在零指代消解方面,大都用到了金标准(gold standard)的句法树信息作为重要特征,如 Zhao 和 Ng[1]、孔芳[2]等人的工作;同时,他们虽然基于同一个机器学习框架来解决零指代问题中所涉及的各个子问题,但是并没有真正意义上的"联合",每个子问题的推断和参数学习均是独立完成的.Iida 等人[3]采用句法树信息作为特征,但他们的工作主要是针对意大利文和日文进行的.此外,尽管他们采用的整数线性规划方法首次将零指代项的识别和零指代项的消解联合在一个框架下进行,但是这种"联合"还仅仅是推断上的"联合",并没有将参数学习也"联合"进来.

本文采用的方法没有使用语料中已标注的正确句法信息作为特征.引入真实的句法树信息在一定程度上利用了真实答案信息,而且在真正的应用中获得正确的句法信息是困难的.因此,本文只使用简单的浅层特征(如词法特征)并研究它们在中文零指代消解中的作用.另一方面,本文采用的方法将零指代项的识别和消解 2 个子问题融合在一个统一的框架下进行联合推断与联合学习.联合推断可以保证对多个子任务所关联的预测变量在一定约束条件下同时进行决策;联合学习可以保证对特征的权重参数进行有效学习.

马尔可夫逻辑网络是组合一阶谓词逻辑与马尔可夫网络的无向图模型.马尔可夫逻辑网络的一阶逻辑规则分为局部规则和全局规则 2 类.局部规则根据多个观察谓词的逻辑组合来对预测谓词的真假进行判断;全局规则用于刻画不同预测谓词间的逻辑关系.本文基于马尔可夫逻辑网络将零指代项的识别与消解 2 个子任务融合在统一的框架下进行联合推断与联合学习.

2.1 局部规则

局部规则即关联观察谓词到一个唯一的预测谓词.对于中文零指代消解问题而言,需要定义一系列的观察谓词来描述单个零指代项的属性或单个候选先行语的属性以及它们之间关系的属性.下面是本文使用的 3 类观察谓词及其解释,其中,n_i 表示语料库中出现的第 i 个零指代项;a 表示零指代项 n_i 附近的位置标记,取值范围为[-3,-1]和[1,3],分别指示 n_i 左侧和右侧最邻近的 3 个词语的位置.

1) 描述候选零指代项 n_i 的谓词

① $zeroNearWord(n_i,a,w)$:在 n_i 附近的位置 a 出现的词为 w;

② $zeroNearPos(n_i,a,p)$:在 n_i 附近的位置 a 出现的词 w 的词性为 p;

③ $zeroNearEntype(n_i,e)$:在 n_i 附近出现实体类型 e;

④ $zeroVerb(n_i,v)$:n_i 直接邻近的动词为 v.

2) 描述候选先行语 m_j 的谓词

① $antNearWord(m_j,a,w)$:在 m_j 附近的位置 a 出现的词为 w;

② $antNearPos(m_j,a,p)$:在 m_j 附近的位置 a 出现的词 w 的词性为 p;

③ $antHasWord(m_j,a,w)$:在 m_j 内部的位置 a 出现的词为 w;

④ $antHasPos(m_j,a,p)$:在 m_j 内部的位置 a 出现的词 w 的词性为 p;

⑤ $antHasEntype(m_j,e)$:在 m_j 内部出现实体类型 e;

⑥ $antHeadWord(m_j,w)$:m_j 的中心词为 w;

⑦ $antHeadPos(m_j,p)$:m_j 的中心词 w 的词性为 p.

3) 描述候选零指代项 n_i 与候选先行语 m_j 之间关系的谓词

① $sentDistance(n_i,m_j,sd)$:n_i 和 m_j 的句子距离为 sd;

② $npDistance(n_i,m_j,nd)$:n_i 和 m_j 的名词短语距离为 nd;

③ $betweenWord(n_i,m_j,a,w)$:在 n_i 和 m_j 之间所在位置为 a 的词为 w;

④ $betweenPos(n_i,m_j,a,p)$:在 n_i 和 m_j 之间所在位置为 a 的词 w 的词性为 p.

由于要解决的中文零指代任务仅有零指代项的识别和零指代项的消解 2 个子任务,因此与之对应的预测谓词仅有如下 2 条:

① $isZero(n_i)$:表示 n_i 是零指代项;

② $zeroResolution(n_i,m_j)$:表示零指代项 n_i 与先行语 m_j 发生指代关系.

需要注意的是,本文没有使用句法信息相关的观察谓词.

与中文零指代问题相关的局部规则有 2 类:1)用于零指代项识别;2)用于零指代项消解.下面是本文使用的 2 类规则及其解释,其中,符号“+”指示对于规则中的变量后带有“+”的所有可能取值进行扩展,每个可能取值都对应一个相应规则的权重.

1) 用于零指代项识别的局部规则

$zeroNearWord(n_i,a,w+)\Rightarrow isZero(n_i)$;

$zeroNearPos(n_i,a,p+)\Rightarrow isZero(n_i)$;

$zeroNearEntype(n_i,e+)\Rightarrow isZero(n_i)$;

$zeroNearWord(n_i,a,wa+)\wedge zeroNearWord(n_i,a+1,wb+)\Rightarrow isZero(n_i)$;

$zeroNearPos(n_i,a,pa+)\wedge zeroNearPos(n_i,a+1,pb+)\Rightarrow isZero(n_i)$;

$zeroNearWord(n_i,a,wa+)\wedge zeroNearPos(n_i,a,pa+)\Rightarrow isZero(n_i)$;

$zeroNearWord(n_i,a,wa+)\wedge zeroNearPos(n_i,a,pa+)\wedge zeroNearWord(n_i,a+1,wb+)\wedge zeroNearPos(n_i,a+1,pb+)\Rightarrow isZero(n_i)$;

$zeroVerbWord(n_i,v+)\Rightarrow isZero(n_i)$.

2) 用于零指代项消解的局部规则

$antHasWord(m_j,b,w+)\vee antHasPos(m_j,b,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHasWord(m_j,b,w+)\wedge antHasPos(m_j,b,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHasEntype(m_j,e+)\Rightarrow zeroResolution(n_i,m_j)$;

$antNearWord(m_j,b,w+)\vee antNearPos(m_j,b,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antNearWord(m_j,b,w+)\wedge antNearPos(m_j,b,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHeadWord(m_j,w+)\vee antHeadPos(m_j,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHeadWord(m_j,w+)\wedge antHeadPos(m_j,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHasWord(m_j,b,wa+)\wedge antHasWord(m_j,b+1,wb+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHasPos(m_j,b,pa+)\wedge antHasPos(m_j,b+1,pb+)\Rightarrow zeroResolution(n_i,m_j)$;

$antHasWord(m_j,b,wa+)\wedge antHasWord(m_j,b+1,wb+)\wedge antHasPos(m_j,b,pa+)\wedge antHasPos(m_j,b+1,pb+)\Rightarrow zeroResolution(n_i,m_j)$;

$sentDistance(n_i,m_j,sd+)\Rightarrow zeroResolution(n_i,m_j)$;

$npDistance(n_i,m_j,nd+)\Rightarrow zeroResolution(n_i,m_j)$;

$betweenWord(n_i,m_j,a,w+)\vee betweenPos(n_i,m_j,a,p+)\Rightarrow zeroResolution(n_i,m_j)$;

$betweenWord(n_i, m_j, a, w+) \wedge betweenPos(n_i, m_j, a, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge npDistance(n_i, m_j, nd+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge antHeadWord(m_j, w+) \wedge antHeadPos(m_j, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge betweenWord(n_i, m_j, a, w+) \wedge betweenPos(n_i, m_j, a, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge antHeadWord(m_j, w+) \wedge antHeadPos(m_j, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge betweenWord(n_i, m_j, a, w+) \wedge betweenPos(n_i, m_j, a, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge antHasWord(m_j, b, w+) \wedge antHasPos(m_j, b, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge antNearWord(m_j, b, w+) \wedge antNearPos(m_j, b, p+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge antNearWord(m_j, b, wb+) \wedge antNearPos(m_j, b, pb+) \wedge zeroNearWord(n_i, a, wa+) \wedge zeroNearPos(n_i, a, pa+) \Rightarrow zeroResolution(n_i, m_j)$;

$zeroVerbWord(n_i, v+) \wedge sentDistance(n_i, m_j, sd+) \wedge npDistance(n_i, m_j, nd+) \wedge antHasWord(m_j, b, wb+) \wedge antHasPos(m_j, b, pb+) \wedge zeroNearWord(n_i, a, wa+) \wedge zeroNearPos(n_i, a, pa+) \Rightarrow zeroResolution(n_i, m_j)$.

2.2 全局规则

中文零指代消解的全局规则用于刻画预测谓词 $isZero(n_i)$ 和 $zeroResolution(n_i, m_j)$ 的关联关系,从而为零指代项的识别和零指代项的消解 2 类预测变量构建相互制约的条件.需要说明的是,这里的全局规则是硬约束,即不像局部规则那样为每个一阶逻辑规则分配一个可以学习的权重系数,而是要绝对满足的.模型所采用的全局规则的一个直观解释是:任何一个被识别为真的零指代项的候选,在文本中其所处位置之前必然存在一个唯一的先行语与其发生指代关系;反过来也一样,即对于任何一个发生零指代消解的先行语,在文本中其所处位置之后必然存在一个或多个零指代项与其发生指代关系.如下是相关的规则:

1) 仅消解零指代项.零指代消解的存在必然导致零指代项的存在.

$$zeroResolution(n_i, m_j) \Rightarrow isZero(n_i), \forall n_i, m_j.$$

2) 零指代项必须被消解.零指代项的存在必然导致相应先行语的存在.

$$isZero(n_i) \Rightarrow |\ \exists m_j : zeroResolution(n_i, m_j)\ | == 1, \forall n_i.$$

3) 不消解非零指代项.非零指代项必然导致不存在相应的先行语.

$$!isZero(n_i) \Rightarrow |\ \exists m_j : zeroResolution(n_i, m_j)\ | == 0, \forall n_i.$$

在具体实现时,以上的 3 类全局规则均可转化为如下所示线性约束条件(其中,变量 y_i 表示谓词 $isZero(n_i)$,变量 $z_{i,j}$ 表示谓词 $zeroResolution(n_i, m_j)$;谓词取真则变量取 1,谓词取假则变量取 0).这样,可以基于整数线性规划来求解,详细内容可参考 Iida 等人[3].

1) 仅消解零指代项.零指代消解的存在必然导致零指代项的存在,即:

$$z_{i,j} \leqslant y_i, \forall i, j.$$

2) 零指代项必须被消解.零指代项的存在必然导致相应先行语的存在,即:

$$y_i = \sum_{j \in Candidate_i} z_{i,j}, \forall i.$$

3) 不消解非零指代项.非零指代项必然导致不存在相应的先行语,即:

$$y_i \geqslant \frac{1}{|\ Candidate_i\ |} \sum_{j \in Candidate_i} z_{i,j}, \forall i.$$

2.3 推断方法

本文采用基于整数线性规划的马尔可夫逻辑网络的推断方法的一个变种:基于割平面 CPI(cutting plane inference)的整数线性规划推断方法[13-14].其目标就是最大化式(1)的一个后验概率.这里用 $\boldsymbol{x}$ 表示所有的观察谓词,用 $\boldsymbol{y}$ 来表示所有的预测谓词.

$s(\boldsymbol{y},\boldsymbol{x})$为将每个特征函数根据特征权重进行线性加权后的打分函数.

$$\hat{\boldsymbol{y}}=\arg\max_{\boldsymbol{y}} p(\boldsymbol{y}\mid\boldsymbol{x})\approx\arg\max_{\boldsymbol{y}} s(\boldsymbol{y},\boldsymbol{x}),\quad(1)$$

其中

$$s(\boldsymbol{y},\boldsymbol{x})=\sum_{(\phi_i,w_i)\in M} w_i\sum_{c\in C^{n\phi_i}} f_c^{\phi_i}(\boldsymbol{y},\boldsymbol{x}).\quad(2)$$

函数 $f_c^{\phi_i}(\boldsymbol{y},\boldsymbol{x})$表示在给定常量 c、观察谓词 $\boldsymbol{x}$ 和预测谓词 $\boldsymbol{y}$ 的情况下,特征函数 ϕ_i 的二元取值函数.同时,每个预测谓词只能取 0 或 1(代表 false 或 true).当推断最佳的 $\boldsymbol{y}$ 时,全局规则应当作为硬约束被满足.所以该推断问题可以很容易地转化为整数线性规划可解的问题.

算法 1. 基于割平面的推断方法.

输入:马尔可夫逻辑网络 L、常量集合 V、基本求解器 BS、初始化闭原子 G^0、观察谓词 $\boldsymbol{x}$、最大迭代次数 $maxIterations$、一阶逻辑规则函数 Φ、一阶逻辑规则系数 $\boldsymbol{w}$;

输出:预测谓词 $\boldsymbol{y}'$.

$i\leftarrow 0$;

$\boldsymbol{y}'\leftarrow\boldsymbol{0}$;

repeat

　$i\leftarrow i+1$;

　$\boldsymbol{y}\leftarrow solve(G^{i-1},\boldsymbol{x})$采用整数线性规划推断方法 Lpsolve① 工具包作为基本求解器 BS;

　if $s(\boldsymbol{y},\boldsymbol{x})>s(\boldsymbol{y}',\boldsymbol{x})$ then

　　$\boldsymbol{y}'\leftarrow\boldsymbol{y}$;

　end if

　for each $(\Phi,\boldsymbol{w})\in L$ do

　　$G_\phi^i\leftarrow G_\phi^{i-1}\bigcup Separate(V,\phi,\boldsymbol{w},\boldsymbol{y},\boldsymbol{x})$;

　end for

until $(G^i=G^{i-1})\vee(i>maxIterations)$;

return $\boldsymbol{y}'$.

算法 1 类似于算法 LazySAT,能够在约束条件较多的时候有效地减少问题的复杂性,可以降低时间复杂度,使问题在可接受的时间范围内得到结果,适用于约束条件复杂时的解码问题.实验中采用 Sebastian Riedel 开发的工具 thebeast(https://code.google.com/p/thebeast).

2.4　参数学习

在参数学习方面,本文采用在线学习算法 MIRA(margin infused relaxation algorithm)进行参数估计.MIRA 算法通过对真实结果和所有错误的结果之间构建一个最大间隔的方式来学习权重参数.这可以通过解一个如下形式的二次规划来实现:

$$\min\ \|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|;\quad(3)$$

$$\text{s.t.}\quad s(\boldsymbol{y}_i,\boldsymbol{x}_i)-s(\boldsymbol{y}',\boldsymbol{x}_i)\geqslant L(\boldsymbol{y}_i,\boldsymbol{y}'),$$

$$\forall\,\boldsymbol{y}'\neq\boldsymbol{y}_i,(\boldsymbol{y}_i,\boldsymbol{x}_i)\in D,$$

其中,$D=\{(\boldsymbol{y}_i,\boldsymbol{x}_i)\}_{i=1}^N$表示 N 个训练实例(每个实例表示数据集中的一个单文档);$\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 分别表示第 i 个训练实例(即一个单文档)中的全部观察谓词和预测谓词;t 表示迭代轮次.在参数学习算法的选择上,本文采用 1-best MIRA.在每次迭代中尝试找到这样的 $\boldsymbol{w}_t$,其可以保证真实的 $\boldsymbol{y}_i$ 与预测的最佳 $\boldsymbol{y}'$(即最高得分的 $s(\boldsymbol{y}',\boldsymbol{x}_i)$等价于 2.3 节提到的 $\hat{\boldsymbol{y}}$)至少要大于等于损失函数值 $L(\boldsymbol{y}_i,\boldsymbol{y}')$,同时保证 $\boldsymbol{w}_{t-1}$ 的改变尽可能小.在损失函数 $L(\boldsymbol{y}_i,\boldsymbol{y}')$的选择上,采用错误预测谓词的闭原子个数来进行度量.全局规则作为硬约束,需要在每次迭代中推断最佳的 $\boldsymbol{y}'$ 时被满足,这样可以使得学到的特征权重更加有效.

3　实　　验

3.1　数据集

本节采用 OntoNotes 3.0 的中文数据集(https://catalog.ldc.upenn.edu/LDC2009T24)来构建实验语料集.该数据包括广播对话(broadcast conversation,BC)、广播新闻(broadcast news,BN)、杂志(magizane,MZ)以及新闻专线(newswire,NW)等内容,不同语料类型的来源如下所示:

1) 广播对话.来源于中国中央电视台(CCTV)、美国有线电视新闻网(CNN)、微软全国广播公司(MSNBC)、凤凰卫视(Phoenix TV).

2) 广播新闻.来源于中央广播电台(CBS)、中华电视公司(CTS)、美国之音(VOA)、中国中央电视台(CCTV)、中央人民广播电台(CNR).

3) 新闻杂志.来源于光华杂志(Sinorama).

4) 新闻专线.来源于新华社(Xinhua News Agency).

由于广播对话比较口语化,在实验中去掉了这部分数据.

OntoNotes 3.0 所标注的中文零指代语料库的零元素类别如下所示:

1) -NONE-*T*:主题化结构或宾语前置结构中的省略语形式.

① http://sourceforge.net/projects/lpsolve/

2) -NONE-* :提升结构和被动结构中的省略语形式.

3) -NONE-* PRO* :控制结构中不能被明显成分替换的省略语形式.

4) -NONE-* pro* :主语或宾语发生缺省的省略语形式.

5) -NONE-* RNR* :发生预指(cataphora)的省略语形式.

为了使研究更加有针对性,本文仅处理-NONE-* pro* 类型的零元素形式,因为这类零元素是真正缺省了主语及宾语的零形式.由于在这类零元素中主语的缺省占极大比例(超过 90%),本文仅处理位于谓语动词之前且在标点符号之后的主语缺省的零元素形式.同时,针对每个真实零元素,令该零元素同与其在一定句子距离(实验中句子距离设为 6)范围内的真实先行语一起构建零指代对正例.

经过如上步骤的处理之后,共得到 319 个零指代对(即零指代项与先行语搭配对)正例,分布于 230 篇文档中.文档中的所有其他候选零指代项和候选先行语的搭配对构成负例.实验中采用五折交叉验证的方式基于这 230 篇文档进行.评价方法将采用信息检索中常用的召回率(R)、准确率(P)以及调和平均值(F).

3.2 实验结果与分析

为了验证联合学习方法的有效性,本文采用 4 种方法进行对比实验.$\text{Baseline}_{\text{maxent}}$ 采用与 2.1 节中所描述的局部规则相同的特征,并基于 Opennlp 工具包(Apache 基金会下面的一个机器学习工具包,用于处理自然语言文本,http://opennlp.apache.org)所实现的最大熵分类器进行训练并测试;MLN_z-Local 系统只使用 2.1 节所描述的局部规则而不包括任何全局规则;MLN_z-Local+Global 系统在参数学习过程中,仍然采用局部规则学习特征权重,但是在推断过程中将全局规则作为约束条件,求解一组使目标函数值最大化的决策变量;MLN_z-Joint 系统以 MLN_z-Local+Global 系统为基础,在能够执行联合推断的基础上,增加了联合学习(即在特征权重的学习过程中同时使用了局部规则与全局规则),使得学习到的特征权重值更加有效.

可以看出,$\text{Baseline}_{\text{maxent}}$ 是基于最大熵分类器的系统;MLN_z-Local 是基于马尔可夫逻辑网络的独立学习系统;MLN_z-Local+Global 是基于马尔可夫逻辑网络的独立学习结合联合推断系统;MLN_z-Joint 则是基于马尔可夫逻辑网络的联合学习结合联合推断系统.实验结果如表 1 所示:

Table 1 Experimental Results of four Systems

表 1 4 种系统的实验结果比较 %

Model	Zero Pronoun Detection			Zero Anaphora Resolution		
	R	P	F	R	P	F
$\text{Baseline}_{\text{maxent}}$	14.9	46.7	22.59	10.5	**50.0**	17.36
MLN_z-Local	16.2	39.1	22.91	11.5	43.2	18.16
MLN_z-Local+Global	15.3	**50.0**	23.43	12.8	38.5	19.21
MLN_z-Joint	**19.0**	33.3	**24.20**	**16.2**	33.3	**21.80**

实验发现,基于马尔可夫逻辑网络的方法及其各种变种有着较快的收敛速度,迭代次数一般设为 10.需要指出的是,为了避免出现数据稀疏问题,针对需要扩展的词形和词性特征只有当其出现次数大于等于 2 时才会被选择到实验中来.

从表 1 可以观察到,$\text{Baseline}_{\text{maxent}}$ 系统的效果最差,在零指代项识别任务上获得的召回率、准确率以及 F 值分别为 14.9%,46.7%,22.59%;在零指代项消解任务上获得的召回率、准确率以及 F 值分别为 10.5%,50.0%,17.36%. MLN_z-Local 系统比 $\text{Baseline}_{\text{maxent}}$ 系统的效果稍好,在零指代项识别任务上获得的召回率、准确率以及 F 值分别为 16.2%,39.1%,22.91%;在零指代项消解任务上获得的召回率、准确率以及 F 值分别为 11.5%,43.2%,18.16%.而联合推断系统 MLN_z-Local+Global 要优于独立学习系统 MLN_z-Local,其在零指代项识别任务上获得的召回率、准确率以及 F 值分别为 15.3%,50.0%,23.43%;在零指代项消解任务上获得的召回率、准确率以及 F 值分别为 12.8%,38.5%,19.21%.效果最好的是联合学习系统 MLN_z-Joint,其在零指代项识别任务上获得的召回率、准确率以及 F 值分别为 19.0%,33.3%,24.2%;在零指代项

消解任务上获得的召回率、准确率以及 F 值分别为16.2%,33.3%,21.8%.联合学习模型在学习特征权重时考虑了相应的硬约束限制条件(即2.2节中介绍的仅消解零指代项、零指代项必须被消解以及不消解非零指代项3个全局规则),因此会比分步的独立学习模型获得更加准确的特征权重,从而使实验结果得到提高.

表1所示的实验结果总体是较低的,最主要的原因是零指代项识别的效果不理想.当给定真实零指代项的情况下(即假定零指代项识别的召回率、准确率和 F 值均为100%时),不同系统基于五折交叉验证的实验结果如表2所示:

Table 2　Experimental Results with Golden Zero Anaphoras

表2　给定真实零指代项情况下的实验结果比较　%

Model	Zero Pronoun Detection			Zero Anaphora Resolution		
	R	P	F	R	P	F
$\text{Baseline}_{\text{maxent}}$	100.00	100.00	100.00	61.70	61.70	61.70
$\text{Baseline}_{\text{rule}}$	100.00	100.00	100.00	63.82	63.82	63.82
MLN_{z}-Gold	100.00	100.00	100.00	**65.96**	**65.96**	**65.96**

表2中增加了 $\text{Baseline}_{\text{rule}}$ 方法,该方法采用简单的基于规则的方法为每个真实零元素选择最近的一个候选先行语作为真实先行语;同时,由于给定了真实零指代项,全局规则在这里不再适用,所以统一用 MLN_{z}-Gold 来表示基于马尔可夫逻辑网络框架的方法.实验结果表明,$\text{Baseline}_{\text{rule}}$ 在不采用其他知识的情况下效果是比较理想的,甚至超出了基于最大熵分类器的方法 $\text{Baseline}_{\text{maxent}}$;$\text{MLN}_{\text{z}}$-Gold 依旧取得了最佳的实验结果;在给定真实零指代项的情况下,3种方法的实验结果均有大幅提高,也证明了零指代项识别任务的提升对于有效解决零指代项消解任务起着重要的作用.

4　小　　结

本文采用马尔可夫逻辑网络框架研究了中文零指代消解问题.基于该框架,中文零指代消解的2个相关子问题:零指代项的识别和零指代项的消解,可以很自然地融合在一个统一的机器学习框架下进行联合推断与联合学习.同时,由于马尔可夫逻辑网络强大的特征表达能力,常规机器学习模型中所使用的特征可以很直观地以局部规则的形式进行描述.而多个子问题的预测变量间的关联关系也很容易用全局规则的形式进行刻画,再结合有效的推断方法和参数学习方法,可以增强模型处理问题的能力.

实验所用数据集来自于 OntoNotes 3.0 的中文宾州树库语料.实验结果显示,基于马尔可夫逻辑网络的联合学习系统要优于独立学习系统以及独立学习结合联合推断的系统,同时也超出了其他 Baseline 系统的性能.这表明,马尔可夫逻辑网络模型适用于中文零指代消解问题.

参　考　文　献

[1] Zhao Shangheng, Ng Hwee Tou. Identification and resolution of Chinese zero pronouns: A machine learning approach [C/OL] //Proc of the 2007 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007: 541-550. [2014-06-28]. http://www.aclweb.org/anthology/D/D07/D07-1057.pdf

[2] Kong Fang, Zhou Guodong. A tree kernel-based unified framework for Chinese zero anaphora resolution [C/OL] //Proc of 2010 Conf on Empirical Methods in Natural Language Processing. 2010: 882-891. [2014-06-28]. http://www.aclweb.org/anthology/D/D10/D10-1086.pdf

[3] Iida R, Poesio M. A cross-lingual ILP solution to zero anaphora resolution [C/OL] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 804-813. [2014-06-28]. http://www.aclweb.org/anthology/P/P11/P11-1081.pdf

[4] Yeh Chinglong, Chen Yichun. Zero anaphora resolution in Chinese with shallow parsing [J]. Journal of Chinese Language and Computing, 2004, 17(1): 41-56

[5] Chen Ping. Chinese discourse analysis for zero anaphora [J]. Zhongguo Yuwen, 1987(5): 363-378 (in Chinese)
(陈平. 汉语零形回指的话语分析[J]. 中国语文, 1987(5): 363-378)

[6] Weng Yiqin. Zero anaphora in Chinese narratives: From a cognitive perspective [D]. Shanghai: Fudan University, 2006 (in Chinese)
(翁依琴. 汉语零形回指的认知研究[D]. 上海: 复旦大学, 2006)

[7] Huang Xian, Zhang Keliang. Zero anaphora in Chinese—The state of art [J]. Journal of Chinese Information Processing, 2009, 23(4): 10-15 (in Chinese)
(黄娴, 张克亮. 汉语零形回指研究综述[J]. 中文信息学报, 2009, 23(4): 10-15)

[8] Grosz B, Joshi A, Weinstein S. Centering: A framework for modelling the local coherence of discourse [J]. Computational Linguistics, 1995, 21(2): 203-225

[9] Converse P S. Pronominal anaphora resolution [D]. Philadelphia, Pennsylvania: University of Pennsylvania, 2006

[10] Hobbs J R. Resolving pronoun references [J]. Lingua, 1978, 44(4): 311-338

[11] Iida R, Inui K, Matsumoto Y. Zero-anaphora resolution by learning rich syntactic pattern features [J]. ACM Trans on Asian Language Information Processing, 2007, 6(4): 1-22

[12] Taira H, Fujita S, Nagata M. A Japanese predicate argument structure analysis using decision lists [C/OL] // Proc of 2008 Conf on Empirical Methods in Natural Language Processing. 2008: 523-532. [2014-06-28]. http://www.aclweb.org/anthology/D/D08/D08-1055.pdf

[13] Riedel S. Improving the accuracy and efficiency of MAP inference for Markov logic [C] //Proc of UAI. Arlington, VA, USA: AUAI Press, 2008: 468-475

[14] Riedel S. Efficient prediction of relational structure and its application to natural language processing [D]. Edinburgh, UK: University of Edinburgh, 2009

Song Yang, born in 1986. PhD. His research interests include natural language processing and machine learning.

Wang Houfeng, born in 1965. Professor and PhD supervisor. Senior member of China Computer Federation. His research interests include natural language processing and machine learning.