计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140807
52(11):2460-2467, 2015

省略识别及恢复联合模型研究

尹庆宇 张伟男 张 宇 刘 挺
(哈尔滨工业大学社会计算与信息检索研究中心 哈尔滨 150001)
(qyyin@ir.hit.edu.cn)

A Joint Model for Ellipsis Identification and Recovery

Yin Qingyu, Zhang Weinan, Zhang Yu, and Liu Ting
(*Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin* 150001)

Abstract An ellipsis is a gap in a sentence due to the pragmatics conventional use of grammar. Ellipsis is a ubiquitous phenomenon in daily conversation, especially in Chinese. A question-answering (QA) system can hardly automatically understand sentences with ellipsis. As a result, the QA system may produce wrong answer and thus cannot naturally interact with humans. Therefore, it is important to recover these ellipses in order to gain a better QA system. To automatically recover these ellipsis elements, we take the recovery system into two parts: zero anaphora identification and zero anaphora resolution. When connecting these two parts together, previous work always models the two steps separately, which suffers the error accumulation problem. In order to deal with this problem, we propose a joint model method that performs the zero anaphora identification and zero anaphora resolution simultaneously in a unified framework. Besides, we focus on Chinese dialogue text, which is collected from the interview of broadcast. The experimental results show that the proposed joint model method outperforms the state-of-the-art methods significantly.

Key words joint model; ellipsis recovery; zero anaphora; anaphora resolution; dialogue

摘 要 省略现象在对话中十分普遍,它的存在导致了语句成分的缺失.问答系统往往不能正确理解这些缺省的表述,这样就会产生错误的问答结果,所以,省略恢复在问答系统中是十分必要的.省略恢复通常分为零代词类别恢复、零代词指代消解2个步骤,已有工作主要是将二者顺序执行,因此会造成错误的累加.为了克服上述问题,提出了1种零代词类别恢复和零代词指代消解联合模型(joint model)的方法,旨在通过联合模型融合省略恢复的2个步骤,进而提高恢复效果.实验结果表明,相比较已有的方法,引入联合模型后,省略恢复的性能得到了显著的提升.

关键词 联合模型;省略恢复;零代词;指代消解;对话

中图法分类号 TP18

在交互式问答中,由于情景、对话者的习惯等原因,省略(ellipsis)现象经常发生,表1列举了一些交互式问答中的省略现象.对于人来说,由于人们彼此之间有着相似的知识背景,所以可以很容易地理解这些存在省略的表述.但如果对话的另一方是无人指导的问答系统,由于缺乏相应的知识背景,问答系统往往不能正确理解这些缺省的表述,这样就会在其答案检索等阶段出现问题,产生错误的问答结果.如果能够将语句中省略的成分恢复,将省略语句变为完整语句,就能够避免这种由于省略现象引起的

收稿日期:2014-09-03;**修回日期**:2015-05-12
基金项目:国家自然科学基金项目(61133012,61202277,61472105)

问答错误,进而提高系统的性能,所以,省略恢复对问答系统而言是十分必要的.

Table 1 Example of the Ellipsis in Chinese Dialogue

表 1 中文对话省略现象示例

Serial Number	Sentences
1	我吃[饭]了.
2	现在,[我]为各位介绍具体情况.
3	[他们]拿碘酒擦了擦,要把这个伤口粘上.
4	[日本]先要解决政治问题.

现阶段,在省略恢复的研究上,已有了很多颇有成效的研究. 如文献[1-3]等利用不同的方法,在Chinese Treebank上实现了零代词类别恢复的目标. 将零代词恢复以后,如果想得到1个完整的句子,还应该将这些零代词恢复成具体的成分,即对这些零代词进行指代消解. 然而,如果简单地将零代词识别和指代消解这2个过程独立地进行并序列化地组合在一起,会产生级联错误.

针对于上述问题,本文将省略恢复分为2个部分:零代词类别恢复和零代词指代消解,并引入了在自然语言处理其他领域(如分词、词性标注等[4])广泛应用的且被证明有效的联合模型(joint model). 该联合模型将这2个部分结合,减少了2个部分的级联错误,进而提升系统的性能. 简单地说,联合模型是1种将几个顺序执行的部分连接在一起的方法,它能利用1个部分的结果对其他部分的结果进行纠正,使它们联合在一起保证整体能够得到最优结果.

本文的贡献在于:

1) 首次在中文对话上做完整的省略恢复尝试. 我们将省略恢复分为零代词类别恢复和零代词指代消解2个部分,并对每个部分做了单独研究.

2) 创新地将联合模型的思想应用在省略恢复系统中,减小了串联省略恢复的2个部分带来的错误累加所产生的影响,通过该方法提升了省略恢复系统的效果.

1 具体研究方法

1.1 基于联合模型的省略恢复系统

我们将省略恢复分为零代词类别恢复和零代词指代消解2个部分. 这2个部分是顺序执行的,即先进行零代词类别的恢复工作,然后根据零代词的恢复结果将其恢复成具体成分,也就是零代词的指代消解工作. 为了避免顺序执行2个部分带来的错误累加,提升系统整体的效果,在将2个部分结合的过程中,我们创新地采用了联合模型的方法.

1.1.1 联合模型介绍

联合模型是1种将多个顺序执行的部分融合在一起的方法,它能够从结合的整体出发,使整体中的每个部分达到更好的效果,进而提升整体的效果. 联合模型的1个典型应用是用其融合分词和词性标注过程. 分词就是将句子 s 分解成词的集合 $W=\{w_1,w_2,\cdots,w_n\}$,其中 $w_1,w_2,\cdots,w_n$ 表示分解的 n 个词. 分词时,确定 W 的过程是使目标函数 $Score_1=f(s,W)$ 达到最大值的过程,即能使 $Score_1$ 的值达到最大的 W 为最后的分词结果. 词性标注是将分好词的1句话 (s,W) 按照 $T=(t_1,t_2,\cdots,t_n)$ 来标注,其中 $t_1,t_2,\cdots,t_n$ 表示 n 个词的词性. 词性标注确定 T 的过程就是使得目标函数 $Score_2=g(s,W,T)$ 达到最大值的过程,即能使 $Score_2$ 的值达到最大的 T 为最后的词性标注结果.

考虑将2个过程合并,1种方法是将2个过程简单地顺序叠加,每个阶段得到的结果只需要保证使各自阶段的函数值达到最大即可. 这种方法的优点是:合并过程简单、直观;但其存在1个明显的弊端,即第1个阶段的错误结果可能会对第2个阶段的结果造成很大影响,从而使最后的结果出现较大的偏差,使用联合模型方法能够避免这种问题. 联合模型方法是将这2个阶段视为1个整体,从整体考虑,通过寻找使分词、词性标注的整体结果达到最优的解来决定分词、词性标注的结果[4-6].

简单地,我们设整体的目标函数 $Score=a_1\times Score_1+a_2\times Score_2$,即 $Score=a_1\times f(s,W)+a_2\times g(s,W,T)$. 这样,就可以将原本的单独考虑各部分结果的情况变成从整体来决定各部分的结果. 也就是说,使用联合模型后,分词和词性标注每一步的结果都是能够使系统整体达到最优. 这样能够有效地避免因为分词错误对最后结果造成的影响,利用2个阶段的互相影响来纠正彼此的错误.

1.1.2 具体方法

零代词类别恢复是预测1个词的前1个位置可能出现的零代词类别. 设该阶段的目标函数为 $f(w,t)$,其表示在词 w 前方出现零代词类别 t 的概率. 恢复零代词类别的过程就是寻找 t,使 $Score_1=f(w,t)$ 达到最大的过程,即使 $Score_1$ 达到最大值的零代词类别 t 就是最后的零代词类别恢复结果.

零代词指代消解是在第 1 个阶段判断出词 w 前存在零代词类别 t 的前提下，为其寻找 1 个指代对象词 c. 设该阶段的目标函数为 $g(w,t,c)$，其表示在词 w 前出现零代词类别 t 的前提下，用词 c 替代零代词类别 t 的概率. 零代词指代消解的过程就是寻找词 t，使 $Score_2 = g(w,t,c)$ 达到最大的过程，即使 $Score_2$ 达到最大值的词 c 为消解结果.

如果将这 2 个部分顺序执行. 对于词 w，首先我们应选出 1 个零代词类别 t；然后，根据这个结果，挑选出 1 个指代对象词 c. 正如 1.1 节所提到，这样可能会造成错误的累加. 因此，我们对于整体使用联合模型方法，令目标分值如下：

$$Score = a_1 \times f(w,t) + a_2 \times g(w,t,c). \quad (1)$$

我们的目标变为寻找零代词类别 t 和替代词 c，使整体 $Score$ 达到最大值. 这样能够从整体出发，达到更优的结果. 训练 a_1 和 a_2 的过程可以采用机器学习支持向量机(support vector machine, SVM)算法，对于输入的每个集合(w,t,c)，将其分为“1”类或“0”类. 我们选取的特征是：补充概率 $f(w,t)$ 和替换概率 $f(w,t,c)$. 在实际应用过程中，对于 1 个新的词对(w,t,c)，将其利用训练好的模型分类后，可得到其分为“1”的概率，将它视为词对的 $Score$ 值即可. 对于词 w，遍历所有可能的词对集合(w,t,c)，选取 $Score$ 最大的 1 个词对，其中的零代词类别 t 和替代词 c 就是最终的结果.

1.2 零代词类别恢复

零代词类别恢复是整个省略恢复的第 1 步，它的效果能够对系统的整体性能造成影响.

1.2.1 数据准备

本节我们选取了 OntoNotes 4.0 的 bc 文件夹下的语料，为其进行标注. 表 2 标明了标注的语料文件名称.

Table 2 OntoNotes Files for Annotation

表 2 标注的 OntoNotes 文件列表

Serial Number	Files
1	phoenix_0000-phonenix_0011
2	msnbc_0000
3	cnn_0000-cnn_0004
4	cctv_0000-cctv_0007
5	p2.5_cmn_0001-p2.5_cmn_0061

在我们的标注过程中，根据需求只针对其中的真实代词做出标注，共计 10 类(具体类别见表 3)，总计标记的句子 2 000 句.

表 3 代词类别、说明及其分布

Table 3 Pronouns Categories, Description and Distribution

Pronoun	Description	Distribution
我	First person singular	0.049
我们	First person plural	0.113
你	Second person singular	0.091
你们	Second person plural	0.003
他	Third person masculine singular	0.107
他们	Third person masculine plural	0.162
她	Third person feminine singular	0.037
她们	Third person feminine plural	0.006
它	Third person inanimate singular	0.404
它们	Third person inanimate plural	0.026

1.2.2 具体方法

首先，利用 berkely parser 工具将分好词、标注好的句子转化为语法树，接下来的工作都集中在这个语法树上；得到语法树后，利用转换工具 Penn2Malt 将语法树转化为依存树，这样就得到了句子的词法信息、句法结构信息和依存信息，特征抽取就是在这个基础上进行的.

利用机器学习算法中的最大熵模型构建分类器，对句子中的词进行预测，预测每个词前边是否出现零代词及零代词的类别. 特征的抽取方式参考了文献[2]的方法，并在其基础上作出了一些改进，抽取了共计 5 类特征.

我们规定，当前的词用 w 表示，w 的前驱词用 h 表示. w 的前驱词的意思是指词 w 在依存树中的前驱节点.

1) 词法特征

词法特征能够从词本身的一些特点上对分类结果作出判断.

① w 本身＋h 本身.

② w 的词性＋h 的词性.

③ w 和 h 之间间隔的词的个数.

④ w 和 h 的距离.

⑤ w 和 h 之间间隔的标点符号的个数.

2) 结构特征

本类特征旨在从句法树(parse tree)中寻找垂直方向上的特性，主要是通过词之间的路径信息来表示的. 假设句法树如图 1 所示，其中词 w 是“为”，h 是“介绍”.

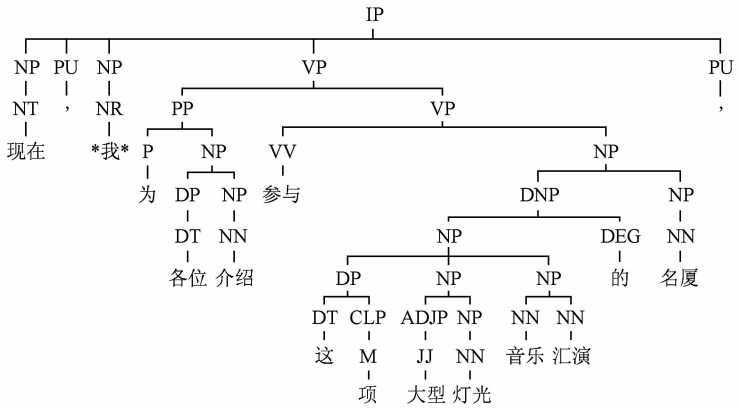


Fig. 1 Example of a parse tree.

图 1 句法树示例

① 词 w 到 w 前的零代词类别前驱的公共父节点路径:本特征是这样抽取的,假设 w 是“为”,则 w 前零代词类别是“我”,这个零代词类别前方出现的词是“现在”(标点忽略),这个词被称为 w 前零代词类别前驱,我们用 A 来表示.从图 1 可以看出,A(“现在”)和 w(“为”)的公共父节点为最上层的 IP 节点,所以在这个例子中抽取出的特征为路径 P_PP_VP_IP.

② 词 w 到 h 的路径:在该句法树中,这个特征是路径 P_PP_NP_NP_NN.

③ 词 h 到 A 的路径:A 的定义同上,所以该路径为 NN_NP_NP_PP_VP_IP_NP_NT.

3) 语法结构特征

语法结构特征主要是利用句法树的一些结构信息来对省略现象作出判断.

① 词 w 是否是句法树的 1 个 IP 节点的第 1 个叶子节点.

② 词 w 是否是 1 个缺少主语或者谓语的 IP 节点的第 1 个叶子节点.

③ w 的左兄弟节点:要求该节点同 w 有共同的 IP 父节点.

④ w 的右兄弟节点:要求该节点同 w 有共同的 IP 父节点.

4) 前驱特征

通过观察发现,对于大多数的省略成分,它们在依存句法中的前驱都是 1 个动词,如果这个动词是 1 个 VP 子树的第 1 个节点,那么这个动词之前很可能存在着省略现象.

① 词 w 是否满足:w 是动词,且 w 是 1 个 VP 节点的第 1 个子节点.

5) 省略信息特征

通过观察数据发现,句子中大部分省略的代词都在这个句子中出现过.比如这样一句话“他走了,关上了门.”在后半句“关上了门”中省略了“他”,而省略的这个词恰好在上半句中出现了.所以我们抽取了这样 1 个特征,来对零代词类别进行预判.

① 句子中,词 w 前是否出现了被省略的代词.

② 词 w 前出现的代词:抽取词 w 前边、同 w 距离最近的代词,这个代词应该是表 3 中 10 种代词类别中的某个.

我们将句子中缺失的零代词类别恢复出来后,下一步的工作就是要将其对应到具体的省略成分上.

1.3 零代词指代消解

零代词指代消解就是将从上一阶段恢复出的零代词类别映射到对话上文中出现的具体成分.零代词类别恢复出来的都是代词,所以我们借鉴了指代消解的相关技术,来进行代词到先行词(省略成分)的消解工作.

1.3.1 数据准备

我们在上一阶段标好的 2 000 句语料的基础上,用 OntoNotes 4.0 的标注规范对额外添加的零代词类别作标注.标注规范如下所示:

我 喜欢 吃〈COREF ID=“1”〉苹果〈/COREF〉.

我 也 喜欢 吃〈COREF ID=“1”〉 * 它 * 〈/COREF〉.

在这个例子中,* 它 * 是上一阶段标注出的零代词类别代词,然后词周围 XML 格式的〈COREF ID=〉表示实体的编号,这里的“它”就可以消解为“苹果”.按照这个规则,对 2 000 句话中的指代现象进行标注,然后数据的预处理工作同上一个阶段相同,将每个句子处理为句法树,并转换为依存树.

1.3.2 具体方法

我们将恢复出的零代词类别和对话上文3句之内出现过的所有名词一一组成词对[8],通过分类器选取1个最佳的词对,用其中的名词替代恢复出的零代词类别.在这里,同样采用机器学习的最大熵模型来构建分类器,将每组词对分为可以消解、不能消解2类.选取的特征如表4所示,对于每个词对(i,j),j表示待消解的代词,i表示替换候选词.

Table 4 Feature Vector

表4 特征向量表

Feature	Description
i_j	Combine word *i* and *j*.
j_num	Whether *j* is singular or plural.
j_sex	Whether *j* is male, female or other.
i_j_sentence_distance	Number of sentences between word *i* and *j*.
i_position	Word *i*'s position in sentence.
j_position	Word *j*'s position in sentence.
i_father	Word *i*'s father node in a parser tree.
j_father	Word *j*'s father node in a parser tree.
ij_head_word	Combine word *i*'s head word and word *j*'s head word.

经过分类,可以将第1阶段恢复出的零代词类别恢复成具体成分;然后利用联合模型的方法,对这2个阶段的结果进行优化.

2 实验结果及分析

2.1 评价方法

采用在自然语言处理领域常用的3个评价指标,对本次实验结果进行评价.即准确率(P)、召回率(R)和F值(F),它们各自的计算公式如下:

$$P=\frac{\text{分类正确的正例数目}}{\text{分类器判为正例的数目}},\tag{2}$$

$$R=\frac{\text{分类正确的数目}}{\text{所有正例的数目}},\tag{3}$$

$$F=\frac{2\times P\times R}{P+R}.\tag{4}$$

2.2 零代词类别恢复结果及分析

在零代词类别恢复的工作中,我们选取了文献[1-2]作为基准,在标注的数据上得到如表5所示的结果.

Table 5 Recovery Results of Zero Pronoun Categories Recovery

表5 零代词类别恢复结果

Approach	*P*	*R*	*F*
Ref[1]	0.300	0.183	0.227
Ref[2]	0.186	0.114	0.141
Ours	**0.348**	**0.230**	**0.276**

从表5不难看出,本文方法是优于基准的.本文方法的优势在于:1)在预测零代词类别的过程中综合了语法、语义等信息,这些信息都能够对预测产生积极的作用;2)充分利用了句子中能够直接对省略零代词类别进行预测的信息,即省略信息这一类特征.我们发现,在标注的数据中,有16%的被省略的零代词在其所在的句子中出现过,所以抽取了这样一类特征,它能够帮助分类器更好地预测零代词的分类.尽管本文方法比基准有一定的提升,但是其效果仍不令人满意,对此我们进行了分析.表3显示了在语料中10个省略类别的分布情况.不难看出,在语料中存在着严重的分布不均问题,有将近一半(40%)的省略类别是"它",而有一些类别例如"她们"只占很小的比例.进一步,我们又对不同的零代词类别的恢复作了统计,如表6所示.通过对结果的观测发现,本文方法对不同类别的零代词的恢复效果和其在语料中出现的频率有关.

Table 6 Recovery Results of Different Zero Pronoun Categories

表6 不同类别零代词的恢复结果

Pronoun	*P*	*R*	*F*
我	0.009	0.002	0.003
我们	0.176	0.036	0.060
你	0.267	0.028	0.050
你们	0.000	0.000	0.000
他	**1.000**	0.003	0.006
他们	0.164	0.018	0.033
她	0.007	0.001	0.002
她们	0.000	0.000	0.000
它	0.362	**0.542**	**0.434**
它们	0.016	0.021	0.018

我们对不同类别零代词恢复的F值和其在语料中的频率作了Pearson相关性统计.结果显示,其相关性达到了0.8,这说明二者有很强的正相关性.因此,可以得出1个结论,数据稀疏问题是导致零代词类别恢复效果不好的主要原因.

2.3 零代词指代消解结果及分析

本节我们选取了文献[9]作为基准,在标注的数据上得到如表7所示的结果:

Table 7 Results of Zero Anaphora Resolution

表7 指代消解结果

Approach	*P*	*R*	*F*
Ref[9]	0.482	0.710	0.574
Ours	**0.548**	**0.881**	**0.676**

从表7可知,本文方法是优于基准的,特别是在召回率上有明显提升.本文方法的优势在于:选取了一些句法结构上的特征,从语法树和依存树上获取了一些候选词和待消解词在结构上的匹配信息,这些信息对消解词的正确选择有一定的帮助,而且本文方法还对待消解的代词做了一些特征上的表示(如其单复数、性别信息等),通过这些表示能够很好地区分不同的代词,这也同样会对消解的结果产生积极的影响.

2.4 联合模型融合结果及分析

我们利用联合模型融合2个步骤后,对新的结果进行重新评价得到如表8所示的结果:

Table 8 Results of Joint Model

表8 联合模型结果

Approach	*P*	*R*	*F*
Zero Pronoun Recovery	0.348	0.230	0.276
Zero Pronoun Recovery+Joint Model	**0.421**	**0.349**	**0.382**
Resolution	0.548	0.881	0.676
Resolution+Joint Model	**0.565**	**0.900**	**0.681**

从表8得到以下结论及分析:

1) 联合模型的方法能够显著提升零代词类别恢复的效果.如2.2节所述,零代词类别恢复效果不好的主要原因是数据稀疏.在此,我们做了另一个实验,对于单类别的零代词类别恢复,即只判断一个词前是否存在零代词类别,其$F=0.8$.由此可以看出,对零代词的具体类别预测得不准确影响了零代词类别恢复的效果.使用联合模型的方法后,通过考虑恢复出的每个代词在其下一个阶段被消解的概率,来纠正在当前阶段出现的代词类别恢复的错误,因此能更为准确地区分可能存在的零代词类别,提升零代词类别恢复效果.

表9是使用联合模型前后不同的零代词类别的恢复结果F值的对比,其中F-old是不使用联合模型的恢复结果F值,F-new是使用联合模型后的恢复结果F值.我们可以看到,对于大部分类别,使用联合模型后其恢复效果都有一定的提升,所以使用联合模型的方法能够纠正在零代词类别恢复阶段产生的错误,且对大部分类别的恢复效果都有提升.

Table 9 Results of Different Zero Pronoun Categories

表9 同零代词类别的结果

Pronouns	*F*-old	*F*-new
我	0.003	0.003
我们	0.060	**0.096**
你	0.050	**0.110**
你们	0.000	0.000
他	0.006	**0.067**
他们	0.033	**0.131**
她	0.002	**0.030**
她们	0.000	0.000
它	0.434	**0.561**
它们	0.018	**0.032**

2) 联合模型方法对指代消解的效果也有一定提升.这是因为在使用联合模型方法融合指代消解和零代词类别恢复的过程中,零代词类别恢复的结果能够对指代消解起到指导作用.在利用联合模型进行指代消解的过程中,我们利用了这2部分的结果的分类信息来训练联合模型的参数值,使正确消解的概率最大化,因此能够在选取消解目标时取得更好的效果.

2.5 整体省略恢复结果及分析

将零代词类别恢复和零代词指代消解2部分结合起来,得到整体的省略恢复结果,如表10所示.其中,单独类别方法是指在识别零代词时不对其具体类别进行区分,统一识别成不区分类别的零代词PRO,然后再进行指代消解;顺序执行方法是指不用联合模型,直接顺序执行省略恢复的2个部分.

Table10 Results of Ellipsis Recovering

表10 省略恢复结果

Approach	*P*	*R*	*F*
Single-Category	0.091	0.121	0.104
Pipeline Method	0.191	0.138	0.160
Joint Model Method	**0.209**	**0.176**	**0.191**

从表10可以得到以下结论:1)识别零代词的具体类别是有必要的.当识别出零代词的具体类别时,其能够在指代消解的过程中起指导作用,使指代消解的效果提升,因此区分零代词的具体类别能够对

省略恢复的整体效果产生积极影响. 2)当我们将 2 个部分线性执行时,由于错误累积会造成省略恢复效果降低,其 F 值只有 0.160;而当我们使用联合模型时,整体的性能有所提高. 这是因为在省略恢复的过程中,联合模型能够从整体的角度出发得到最优解,对前后 2 个部分的结果起到纠正的作用. 也就是说,使用联合模型后能够减少各部分的错误累积,进而提高整体的效果. 因此在最后的结果上,使用联合模型的省略恢复系统明显优于其他系统.

3 相关工作

关于省略恢复的研究主要有零代词类别恢复和零代词指代消解. 零代词类别恢复是省略恢复中十分重要的环节,它的效果能够很大地影响省略恢复整体系统的性能;零代词指代消解是 1 种特殊的省略恢复,它是指利用上下文信息将句子中缺失的指代词恢复成具体成分.

3.1 零代词类别恢复

在最近的一些研究中,利用机器学习的方法,在中文对话的零代词类别恢复上,研究者们已经获得了一些成果. 文献[1]提出了 1 种学习结构的方法来预测零代词类别,并将预测结果应用到机器翻译中,使机器翻译系统的性能得到了提升. 文献[2]在研究中创新地使用了一些新类别的特征,他们首先将句法树转换为依存树,然后从中提取了一些特征,这些特征在预测零代词类别上效果提升明显. 文献[3]提出了 1 种将零代词类别恢复视为打标签问题的方法,利用词法和语法 2 类特征为句子中的每个词来打标签,以判别该词前面是否出现省略现象. 在训练和测试的过程中,如果将每个词打上了“EC”(有省略)或“NEC”(无省略)标签,就相当于构建 1 个二元分类器,识别出句子中是否存在省略现象;但是,这个方法缺少对句子结构信息的利用,因此在多类别标签上的效果并不理想. 文献[10]提出了 1 种有效的零代词类别恢复方法,它融合了词法语法和句法结构等信息,构建了 1 种新的为零代词类别及其位置编码的方式,将零代词类别信息和位置信息编码到句法树的非叶子节点上,这种新的编码方式十分有效,恢复效果显著.

3.2 零代词指代消解

现阶段,在中文对话系统中,有关零指代消解的工作并不多,其原因是中文本身的复杂性和独特性大大增加了该任务的难度,因此很少有显著的研究成果. 零代词指代消解是指在省略指代词的句子中找出缺失指代词位置,并将其补充成指代先行词的过程. 这里,缺少的指代词被称为零代词.

在过去的很长一段时间内,中文零指代消解的方法都局限于基于规则的方法. 文献[11]提出了利用中心理论,将手工标注规则的方法应用到中文零指代消解的研究,取得了一定的效果. 文献[12]提出利用 Hobbs 算法,通过句法分析树完成了零指代消解的工作,也取得了不错的效果. 但是,随着语料规模的不断加大,基于规则的方法遇到了瓶颈,研究者将工作重点转移到了机器学习的方法. 文献[13]提出了 1 种机器学习的方法,为基于机器学习的中文零指代消解方法开创了先河. 他们选取了一系列特征,并利用决策树分类器对中文零指代现象进行判别. 这个方法在中文零指代上的影响很大,尽管该方法在特征选取方面有其局限性,但是不可否认其是 1 个很经典的方法. 后续又有一些研究者在该方法的基础上改进了特征的选取,都获得了较好的结果. 文献[14]提出了 1 种基于树核的中文零指代消解的方法,其在文献[13]的基础上将零代词与其先行词进行了一一对应,将零代词与其可能的先行词候选词一一组合,提取特征进行二元分类(SVM),也很好地对零指代缺失的成分进行了恢复.

4 结论及未来工作

本文提出了 1 种中文对话语料的零代词类别恢复和零代词指代消解的联合模型方法,通过该方法提升了省略恢复系统的效果. 在我们的研究中,对这 2 部分方法单独的优化还不够,因此在今后的工作中我们会加强对各模块方法的单独研究,以获得更好的省略恢复效果.

参 考 文 献

[1] Xiang B, Luo X, Zhou B. Enlisting the ghost: Modeling empty categories for machine translation [G] //Association for Computational Linguistics. Stroudsburg, PA: ACL, 2013: 822-831

[2] Xue N, Yang Y. Dependency-based empty category detection via phrase structure trees [C] //Proc of North American Chapter of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg, PA: ACL, 2013: 1051-1060

[3] Yang Y, Xue N. Chasing the ghost: Recovering empty categories in the Chinese Treebank [C] //Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2010: 1382-1390

[4] Li Z, Zhang M, Che W, et al. Joint models for Chinese POS tagging and dependency parsing [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011: 1180-1191

[5] Bohnet B, Nivre J. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing [C] //Proc of the 2012 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: ACL, 2012: 1455-1465

[6] Wulfsohn M S, Tsiatis A A. A joint model for survival and longitudinal data measured with error [J]. Biometrics, 1997, 57(1): 330-339

[7] Kong F, Zhou G. A clause-level hybrid approach to Chinese empty element recovery [C] //Proc of the 32nd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2013: 2113-2119

[8] Yang X, Su J, Tan C L. Kernel-based pronoun resolution with structured syntactic knowledge [C] //Proc of the 21st Int Conf on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2006: 41-48

[9] Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases [J]. Computational Linguistics, 2001, 27(4): 521-544

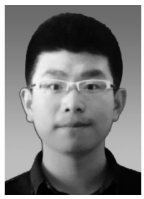
[10] Cai S, Chiang D, Goldberg Y. Language-independent parsing with empty elements [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 212-216

[11] Yeh C L, Chen Y C. Zero anaphora resolution in Chinese with shallow parsing [J]. Journal of Chinese Language and Computing, 2007, 17(1): 41-56

[12] Converse S P. Pronominal anaphora resolution in Chinese [D]. Philadelphia, PA: University of Pennsylvania, 2006

[13] Zhao S, Ng H T. Identification and resolution of Chinese zero pronouns: A machine learning approach [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2007: 541-550

[14] Kong F, Zhou G. A tree kernel-based unified framework for Chinese zero anaphora resolution [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2010: 882-891



Yin Qingyu, born in 1990. PhD, lecturer. His main research interests include ellipsis recovery, information retrieval and question answering.

Zhang Weinan, born in 1985. PhD, lecturer. His main research interests include social computing, information retrieval and question answering (wnzhang@ir.hit.edu.cn).

Zhang Yu, born in 1972. PhD, professor, master supervisor. His main research interests include question answering and personalized information retrieval (zhangyu@ir.hit.edu.cn).

Liu Ting, born in 1972. PhD, professor, PhD supervisor. His main research interests include social computing, information retrieval and natural language processing (tliu@ir.hit.edu.cn).