第 23 卷 第 3 期
2009 年 5 月

中文信息学报
JOURNAL OF CHINESE INFORMATION PROCESSING

Vol. 23, No. 3
May, 2009

文章编号: 1003-0077(2009)03-0003-07

集成多种背景语义知识的共指消解

郎 君,忻 舟,秦 兵,刘 挺,李 生

(哈尔滨工业大学 信息检索研究室,黑龙江 哈尔滨 150001)

摘 要: 共指消解是信息抽取中一个重要子任务。近年来,许多学者尝试利用统计机器学习的方法来进行共指消解并取得了一定的进展。背景知识作为新的研究热点已经被越来越多地利用在自然语言处理的各个领域。该文集成多种背景语义知识作为基于二元分类的共指消解框架的特征,分别在 WordNet、维基百科上提取背景知识,同时利用句子中的浅层语义关系、常见文本模式以及待消解词上下文文本特征。并利用特征选择算法自动选择最优的特征组合,同时对比同样的特征下最大熵模型与支持向量机模型的表现。在 ACE 数据集上实验结果表明,通过集成各种经过特征选择后的背景语义知识,共指消解的结果有进一步提高。

关键词: 计算机应用;中文信息处理;共指消解;背景语义知识;WordNet;维基百科

中图分类号: TP391 **文献标识码**: A

Coreference Resolution with Integrated Multiple Background Semantic Knowledge

LANG Jun, XIN Zhou, QIN Bing, LIU Ting, LI Sheng

(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin, HeiLongjiang 150001, China)

Abstract: The coreference resolution is an important subtask of information extraction. Recently statistical machine learning methods have been substantially attempted for this issue with some achievements. In this paper, we try to integrate the background semantic knowledge, which is a new subject being introduced in every field of NLP nowadays, into the classical pairwise classification framework for coreference resolution. We extract background knowledge from WordNet and Wikipedia, and exploit the semantic role labeling, general pattern knowledge and the context of mention as well. In the experiment, the feature selection algorithm is employed to decide the best features set, on which the maximum entropy model and SVM model are compared for their performance. The experimental results on ACE dataset exhibit the improvement of coreference resolution after adding selected background semantic knowledge.

Key words: computer application; Chinese information processing; coreference resolution; background knowledge; WordNet; wikipedia

1 引言

共指消解就是将篇章内的所有表述划分为现实世界中不同实体等价描述的过程,主要包含人称代词消解和名词短语消解[1]。该问题一直是信息抽取中的重要子任务之一。随着对问题的研究深入,越来越多的研究人员意识到共指消解是人工智能中最难的问题之一,因为共指消解不仅需要语言学方面的知识,例如浅层的词汇、句法知识,还需要较为宏观的篇章和语义知识。最为困难的是,很多时候共指消解需要丰富的背景知识才能完成。

最近十几年,随着消息理解会议(Message Understanding Conference, MUC)以及自动内容抽取(Automatic Content Extraction, ACE)等系列大型国际评测的不断开展,基于统计机器学习的共指消

收稿日期: 2008-08-26 **定稿日期**: 2008-11-01

基金项目: 国家自然科学基金资助项目(60575042, 60503072);国家 863 计划资助项目(2006AA01Z145)

作者简介: 郎君(1981—),男,博士生,主要研究方向是自然语言处理;忻舟(1986—),男,本科生,研究方向为自然语言处理;秦兵(1968—),女,博士,教授,研究方向为自然语言处理、信息检索。

解方法取得了长足的进步。但是,这些方法主要采用的都是一些较为浅层的特征,例如实体之间的距离、性别、单复数、人称、实体类型、字符串匹配、同位、别名等。

近年来,最新的研究都主要集中在如何深入发掘和利用各种语义和背景知识上。

Ponzetto 和 Strube 利用挖掘浅层语义角色(Shallow Semantic Role)、WordNet 和维基百科(Wikipedia)来增强共指消解,在传统共指消解特征框架下主要增加了实体的语义角色、实体对分别在 WordNet 和维基百科上的语义相似度等三种特征[2]。

Bean 和 Riloff 利用共指关系明显的共指实体对在两个领域的语料中进行模板挖掘,并人工对一部分模板进行了删选和增强,利用 Dempster-Shafer 决策模型来做信息融合从而进行共指消解,结果表明不同的代词需要用不同的特征来消解,确定性名词短语偏重使用词汇特征,代词偏重使用上下文语义特征[3]。

Yang 和 Su 利用模板来获得指代词和先行词之间的语义关系,在维基百科上自动挖掘模板,并对模板进行评价和打分,然后利用模板来获得待消解的共指实体对的语义特征[4]。

现在能够利用的背景语义知识来源主要有 WordNet、维基百科、浅层语义角色标注(Semantic Role Labeling, SRL)、上下文模板等。本文主要将这些背景语义知识综合起来考察对共指消解系统的影响,同时结合共指消解问题的特殊性,提出了上下文特征。另外对常规特征在内的特征集合采用自动特征选择的方法整体分析了各种背景语义特征组合对共指消解的作用。结果表明,共指消解在背景语义的支持下能够取得较大的提高,特征选择对于共指消解也是必不可少的环节。

本文按照如下方式组织:第二部分概述了共指消解研究的发展历程以及经典的基于二元分类框架以及用于共指消解的常规特征;第三部分详细介绍了各种背景语义特征的特点和构造方法,并结合代词类共指消解的特点提出了上下文特征;第四部分说明融合各种背景语义知识进行共指消解的系统框架以及自动特征选择的方法;第五部分介绍详细的实验设计以及结果分析;最后是总结和未来工作的展望。

2 相关研究工作概述

共指消解研究的研究可以分为三个阶段[1]:(1) 1978 年～1995 年,以句法分析为基础的基于语言学方法的共指消解,代表方法是 Hobbs 算法以及中心理论;(2) 1995 年～2002 年,这段时间主要是各种基于二元对的分类方法以及基于向量相似度的聚类方法;(3) 2002 年至今,经过上一个阶段的发展,越来越多的研究人员开始考虑如何引入背景知识以及语义知识,同时采用一些全局考虑篇章信息的方法来实现最优化的篇章共指消解。

随着 McCarthy 和 Lehnert 首次将共指消解问题视为二元分类并采用决策树(Decision Trees) C4.5 算法[5]以来,共指消解开始在二元分类的框架下获得了长足的发展。经典的基于二元分类的共指消解系统框架如图 1 所示[1]。

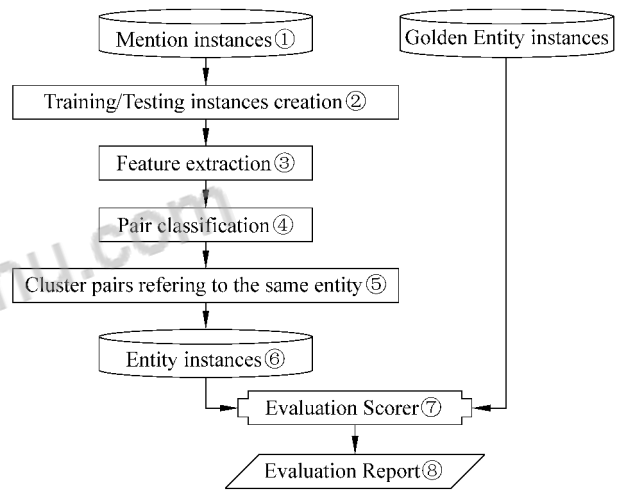

图 1 基于二元分类的共指消解经典框架

图中①表示共指消解处理的对象。一般而言,共指消解系统的输入是预处理中获得的各种实体表述(Mention)。相关预处理主要包括断句、词性标注、命名实体识别、嵌套名词短语识别等。这些前处理一般采用一些相关的模块来获得。共指消解的国际评测中,为了更加精准的评测共指消解算法的性能,主办方一般都会提供标注好 Mention 的语料。②表示从训练语料或者测试语料中构建用于分类器的输入实例。针对训练和测试分别采用不同的实例构建方法。③表示特征抽取。事实上,在二元分类框架下,如何设计需要选定的特征,对于最终的共指消解性能具有决定性的影响。本文的各种背景语义特征在共指消解上的应用就主要体现在这个环节。图中④表示二元分类的机器学习算法。到目前为止,用于共指消解二元分类的机器学习方法主要有贝叶斯(Naïve Bayes)、决策树、支持向量机(Support Vector Machine, SVM)、最大熵(Maximum

Entropy)、条件随机域(Conditional Random Field, CRF)、遗传算法(Genetic Algorithm, GA)、互训练(Co-Training)等。这些方法的一个共同点是都在各种相关特征构成的特征向量的基础上训练得到各种特征的权值或者优选性(主要是决策树能得到优选性)。图中⑤表示 Mention 二元分类的结果合并为 Entity。随后进行的就是共指消解结果的评价(图中⑦所示)以及得到最终的实验结果(图中⑧所示)。

对于两个实体表述(Mention) I 和 J, I 在 J 之前, I 称为先行词, J 称为指代语。

二元分类框架下,共指消解需要的传统特征可以分成四类,即词汇特征(STRING_MATCH, ALIAS),语法特征(I_PRONOUN, J_PRONOUN, J_DEF, J_DEM, NUMBER, GENDER, PROPER_NAME, APPOSITIVE),语义特征(WN_CLASS),距离特征(DISTANCE)[6],如表 1 所示。

表 1　共指消解所用常规特征

名　称	返回类型	说　明
STRING_MATCH	T, F	I、J 分别去掉冠词(a, an, the)和指示代词(this, these, that, those)后进行匹配。匹配成功返回 T,否则返回 F。
ALIAS	T, F	如果 I 是 J 的别名或者 J 是 I 的别名,返回 T,否则返回 F。
I_PRONOUN	T, F	如果 I 是代词(包括反身代词、人称代词和物主代词),返回 T,否则返回 F。
J_PRONOUN	T, F	如果 J 是代词,返回 T,否则返回 F。
J_DEF	T, F	如果 J 是限定性名词短语(定冠词开头),返回 T,否则返回 F。
J_DEM	T, F	如果 J 是指示性名词短语(以 this, that, these, those 开头),返回 T,否则返回 F。
NUMBER	T, F	如果 I 和 J 单复数相同则返回 T,否则返回 F。
GENDER	U, T, F	如果 I 和 J 中存在一个性别不确定,返回 U,如果 I 和 J 性别(包括 male, temale, neutral)相同,返回 T,否则返回 F。
PROPER_NAME	T, F	如果 I 和 J 都是专有名词(主要单词都是首字母大写),返回 T,否则返回 F。
APPOSITIVE	T, F	判断 I 和 J 是否同格, I 和 J 中必须至少有一个是专有名词。通过判断 I 和 J 之间是否有动词或特定标点(如逗号)来判断是否 APPOSITIVE。如"Bill Gates, the chairman of Microsoft Corp."中 Bill Gates 和 Chairman of Mircosoft Corp。
WN_CLASS	U, T, F	如果 I 和 J 有一个不在 WordNet 中则返回 U,如果 I 和 J 是同一个 WordNet 语义类别,返回 T,否则返回 F。
DISTANCE	0, 1, 2, 3	返回 I 和 J 之间相隔的句子数量,句子间隔大于或者等于 3 时,都返回 3。

3　多种背景语义知识

共指消解如同其他自然语言处理问题一样,是一个强不适定问题(Stringly Ill-posed Problems),只有通过提供大量丰富的"约束"(包括知识、经验等),才能使之成为适定性的、可解的问题[7]。共指消解需要采用大量的约束才能解决,而对于二元分类等具体的框架,添加约束的方法就是采用更多合理的特征。本文整合了各种相关的背景语义知识,并提出了结合维基百科的上下文特征。

3.1　基于 WordNet 的背景语义特征

传统的共指消解中 WordNet 仅被用来判断两个名词在 WordNet 上所属的语义类别是否一致,如表 1 中的 WN_CLASS。但是 WordNet 的语义类别由于存在多义、覆盖率等问题,造成了判断语义类别时含有非常多的噪音,影响了结果的准确率。

WordNet 中存在丰富的语义关系,如同义、上下位、整体部分等。本文主要通过计算实体二元对的核心词之间的相似度来利用 WordNet 语义知识。WordNet 相似度有多种计算方法,主要分为两大类,一类是三种基于路径长度的相似度(Rada; Wu & Palmer; Leacock & Chodorow),另一类是三种基于信息内容的相似度(Resnik; Jiang & Conrath; Lin)。本文采用 WN-Similarity 工具包①,该工具包

① http://sourceforge.net/projects/wn-similarity

提供了以上6种相似度的计算方法。

在每一种相似度计算方法中,为了覆盖所有义项,对每种可能的词义都计算一遍相似度(如果I有N种义项,J有M种义项,就有$N \times M$个相似度)。对Mention对〈I,J〉基于WordNet的背景语义特征相似度值如下[2]:

WN_SIMILARITY_BEST:实体二元对$N \times M$个义项对相似度中的最大值。

WN_SIMILARIRY_AVG:实体二元对$N \times M$个义项对相似度的平均值。

由于每种相似度的计算方法都可以分为以上两个特征,所以WordNet相似度共有12个特征。

3.2 基于维基百科的背景语义特征

维基百科是世界各地的志愿者贡献的巨大知识库①。每篇文章都包含了一个实体或一个概念的信息,收集了关于这个实体的特定信息,对于有歧义的页面提供整体部分关系的信息。截至2008年7月,维基百科包含了将近240万个英文页面。

如果Mention是命名实体,检索完整的字符串(Extent),否则检索对应的核心词(Head)。跟踪检索到的所有重定向页面。如果检索到一个有歧义的页面,首先把这个页面中所有的链接都记录下来。如果一个链接含有另一个实体(比如President在Lincoln的页面中),那么就返回这个超链接(例如检索Lincoln返回President of the United States),否则返回有歧义的页面中第一个页面。给定一个候选共指对〈I,J〉,分别检索对应的Wikipedia的页面P_I,P_J,抽取如下的信息作为特征[2]:

I/J_GLOSS_CONTAINS:如果没有对应的维基百科词条,返回U;如果某个实体对应页面的第一段包含另一个实体,则返回T,否则返回F。

I/J_RELATED_CONTAINS:如果没有对应的维基百科词条,返回U;如果某个实体对应页面的某个超链接中包含另一个实体的标题,返回T,否则返回F。

I/J_CATEGORIES_CONTAINS:如果没有对应的维基百科词条,返回U;如果一个实体的某个类别包含另一个实体,则返回T,否则返回F。

I/J_GLOSS_OVERLAP:计算P_I和P_J的第一段的重叠得分,即是P_I和P_J的第一段中重复的单词的个数的平方。

3.3 浅层语义知识

对语料进行浅层语义标注②,可以得到各种语义角色信息。例如"He talked for about 20 minutes"可以被标注为"[ARG0 He] [TARGET talked] [ARGM-TMP for about 20 minutes]"。对标注结果中ARG部分的词和需要判定共指的实体的核心词进行匹配,只要能够匹配上就认为该Mention是该ARG。把ARG与TARGET对作为一个标签加入到特征中,故一个句子中提取出来的标签具有这样的形式ARG_1-$PRED_1$,…,ARG_n-$PRED_n$。对于某个特定的实体,找到它所在句子中的对应的ARG,并返回与之相关的标签。

例如上述例子中,He是一个实体,ARG_0-talked是He的SRL特征。需要注意的时,由于Mention的核心词经常不能和标注结果完全匹配,这里规定只要核心词含在某个ARG的覆盖范围内,就认为该实体的ARG信息就是这个文本。

这里将实体对I,J的浅层语义标签作为特征[2]:

I_SEMROLE:I的语义标签。

J_SEMROLE:J的语义标签。

3.4 共指模板特征

语义相关度对于共指消解任务来说是一个非常重要的因素。为了获得这种语义信息,基于语料的方法通常是抽取某些能表达特定语义关系的模式。这种模式,也称作模板。

本文采用自动利用"种子"挖掘模板的方法来自动生成模板。

首先在语料中抽取I,J之间的文本。人工去掉一些噪音符号,然后将剩下的部分作为标准。例如"strawberry bud weevil, also called strawberry clipper weevil"中,"strawberry bud weevil"和"strawberry clipper weevil"是共指对,可以把", also called"这个模板挖掘出来。但是由于语料中存在一些噪声,比如在维基百科上的文本中进行挖掘时,常会把"||"也作为模板,而这种模板在实际句子中几乎是遇不到,所以我们需要人工去掉它们。详细的模板挖掘方法请见文献[4]。

模板特征如下:

PATTEN:Mention对〈I,J〉是否符合模板,如果符合返回T,否则返回F。

① http://www.wikipedia.org/。

② 本文采用Assert工具进行自动SRL标注,http://cemantix.org/assert。

3.5　上下文特征

以往的特征是针对 Mention 本身的,这样的特征在 Mention 是代词时作用就不明显了。根据统计结果,代词的消解错误率要比名词和名词短语高。因为代词除了指明性别和句法功能外,不能提供额外的信息。根据代词的特点,本文提出一种利用 Mention 周边词作为背景知识的方法,从而就避免了直接从 Mention 找特征。

对于先行词 I,在维基百科中寻找对应页面,把这个页面与指代词 J 周围的词求相似度。比如"That court has now put on hold its ruling banning the pledge of allegiance in public classrooms."该句中"court"和"its"是共指对,"its"周围有"rule","ban","pledge"等词,通过维基百科查询"court"的词条,然后在"count"的词条正文中寻找"rule","ban","pledge"等词,并且记录这些词出现的个数。如果在"count"的词条中经常出现这些词的话,就可以认为"count"与"its"的关系非常密切,从而有可能指向同一个实体。上下文特征如下:

I_CONTEXT: 利用维基百科查询指代词 J 的背景知识页面,然后取先行词 I 周围的几个词(设定为前面 3 个词,后面 5 个词)作为上下文,计算背景知识和上下文之间的相似度。为了易于实现,本文直接判断背景知识里是否含有这些词中的一个或多个(这里的背景知识去掉那些常见的介词、连词、形容词、冠词等),下同。

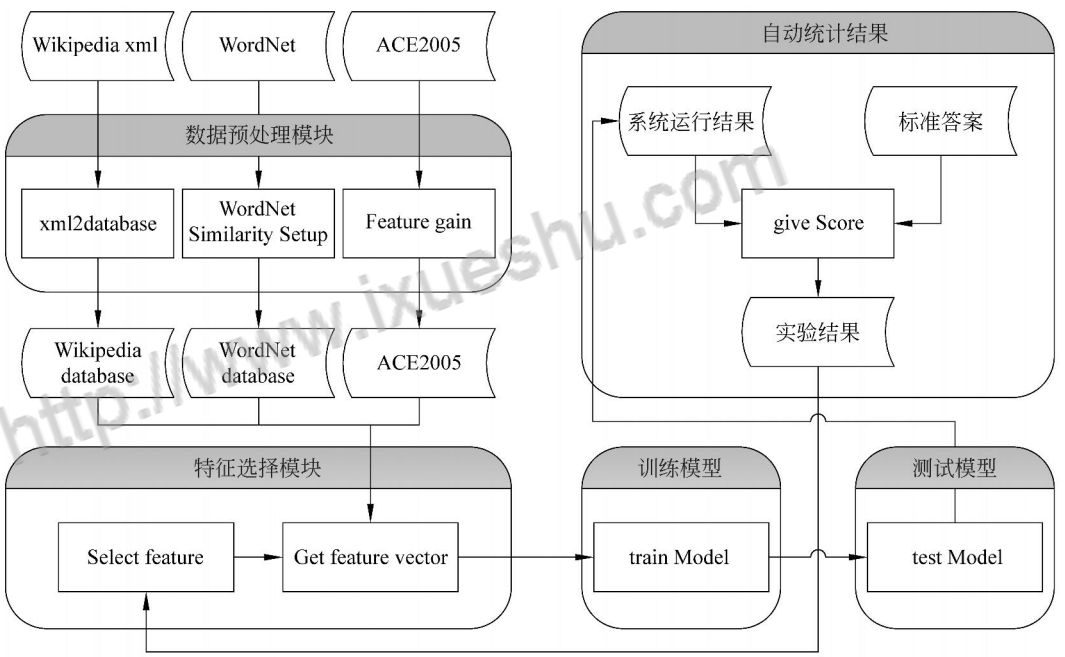

图 2　系统框架图

J_CONTEXT: 利用维基百科查询先行词 I 的背景知识页面,然后取指代语 J 周围的几个词(设定为前面 3 个词,后面 5 个词)作为上下文,计算背景知识和上下文之间的相似度。

4　集成背景语义知识的共指消解

在经典的二元分类共指消解框架下,可以实现传统共指消解特征和上面提出的各种背景语义特征的集成。下面先介绍特征选择算法,然后是整个实验系统的框架。

4.1　共指特征选择算法

面向特征的机器学习框架在进行学习时容易出现的一个问题就是训练不够充分,尤其是在不断加入新特征的情况下。本文的工作就是在原有共指消解常规特征的基础上,加入了多种背景语义特征。在训练集确定不变的情况下,为了增加学习模型的泛化能力,需要采用特征选择方法来筛选有效的特征。

Kohavi 和 John 提出过一种类似于"去皮"的方法,就是每次去掉一个特征以帮助确定哪个特征对系统"贡献"最小。这种方法在特征空间上进行爬山

搜索,开始使用所有特征训练,然后依次去掉一个特征,重新训练。每次去掉表现最坏的一个特征,也就是系统的 F 值上升最高的那个。重复这样的过程,直到结果不在提高为止[6]。这种方法被许多研究者用于特征选择并取得了很好的效果。

4.2 共指消解系统设计

多种背景语义知识的利用,以及相关特征选择算法的采用,使得搭建的共指消解系统更加复杂。本文中使用的系统框架如图 2 所示。实验数据采用 ACE2005。

数据预处理模块:把维基百科的 xml 数据①转换为数据库以便于储存和处理;把 WordNet 的原始数据②进行预处理,用于获得 WordNet 相似度;对 ACE 语料进行预处理,以适合实验需要。

特征选择模块:分为两部分,获得特征向量(Get Feature Vector)部分对给定的语料进行特征识别,返回一个特征向量。特征选择(Select Feature)部分对于上一轮的测试结果,选择最佳的特征继续进行测试和训练,直到结果没有再提高为止。

训练模型:调用特征识别模块对语料进行处理,训练模型。

测试模型:调用特征识别模块对语料进行处理,测试模型,并输出标准答案和系统输出答案。

自动统计结果:对答案进行自动统计并输出结果。

5 实验设计及结果分析

5.1 实验设计

本文使用 ACE2005 的 224 篇 bnews 语料,按 4∶3∶3 分成三部分。第一部分是训练集,用作训练模型;第二部分是开发集,用来测试训练出来的模型,并为重新选择特征提供数据;第三部分是测试集,对于经过特征选择后的最优模型进行测试,得到实验结果。

使用 Soon 等的方法[8]选用所有相邻的共指对作为正例,如共指链 A1-A2-A3-A4,选择 A1-A2,A2-A3,A3-A4 作为正例。反例的选取是在相邻两个共指对之间找到所有与之不共指的 Mention,把它与指代语配对作为反例。例如 A1 和 A2 之间有 a,b,B1,其中 A1,A2 是一对共指对,a,b 是独立的 Mention,B1 是其他共指链中的 Mention。那么 a-A2,b-A2,B1-A2 都作为反例。

在测试时,在一篇文档内从前往后扫描,对每个 Mention,都寻找它之前的所有 Mention,构成 Mention 对用于判断,直到遇到第一个被系统判断为正例的 Mention 为止或者遇到开头。

实验选定最大熵③和 SVM④两种机器学习方法用于实验。由于最大熵的训练速度比 SVM 快,在进行特征选择时采用最大熵,确定最优的特征组合后使用 SVM 进行重新训练和测试。由此也能对比两种机器学习方法在最终选定的最优特征集合上的性能差异。模型输出结果的阈值设为 0.5,即当概率超过 50%时,系统就认为这两个 Mention 共指。

系统的评价方法采用目前国际上普遍采用 CEAF(Constrained Entity-Aligned F-Measure)[9]。

5.2 实验结果及分析

系统在基本特征及其他各种背景特征下的结果如表 2 所示。结果表明,在基本特征上各种背景知识单独使用,都能在一定程度上提高结果,特别是在召回率上。需要说明的是,这里单独的背景知识在模型中还被分为了诸多子特征,表中的每种背景知识都是使用全部的子特征,所以受某些子特征影响,整体结果可能会降低。而特征选择算法是把所有的子特征混合在一起进行筛选,最终获得最优的特征组合。

表 2 中结果也符合常识,即加入背景知识后,系统的召回率有所提高。显然,当系统获得了更多的背景知识后,它就能够识别出更多的共指关系,使原本不能被自动识别为共指的共指对也获得识别。因为经典的二元分类框架是自底向上的一种链接合并过程,加入更多能被正确识别的共指关系,就能提高所有 Mention 的合并程度,从而实现召回率的提高。当然,部分由此错误添加的共指关系也会导致系统的精确率下降。但是最终评价系统整体性能的 F 值得到了 5%的提高。

① 使用的维基百科是 20071018 日备份的,共有 5 836 166 个页面,包括 Articles, templates, image descriptions, and primary meta-pages。

② WordNet2.1, http://wordnet.princeton.edu/2.1/WordNet-2.1.exe。

③ http://homepages.inf.ed.ac.uk/s0450736/maxent.html version1.1

④ http://www.cs.cornell.edu/people/tj/svm_light/ version 6.01

表 2　加入特征及实验结果

特　征	召回率 / %	精确率 / %	F 值 / %
12 项基本特征(Base)	76.90	54.77	63.98
Base + 12 项 WordNet 特征	78.99	53.91	64.09
Base + 4 项维基百科特征	76.36	52.23	62.03
Base + 2 项 SRL 特征	76.77	54.47	63.73
Base + 1 项模板特征	76.93	54.88	64.06
Base + 2 项上下文特征	75.84	55.56	64.14
全部特征(33 项)	79.35	50.42	61.66
特征选择后的优化特征(27 项)	79.10	60.53	68.58
SVM + 优化特征(27 项)	82.97	56.96	67.55

在所有的特征中,上下文特征对精确率的提高最大(2 %)。因为上下文特征由于不考虑词本身的内容,而是看这个词周围的上下文,对于代词这类信息量非常少的 Mention,上下文特征发挥了很大的作用,而这个长处就在精确率上有所体现了。

另外,从表 2 最后两行可以看出,在同样的特征集合上最大熵模型在精确率上高于 SVM,SVM 在召回率上高于最大熵,综合看来最大熵方法比 SVM 模型的结果略好。这说明机器学习模型的选用对最终系统的性能有很小的影响。

下面分析特征选择过程,如表 3 所示。

表 3　特征选择过程

特　征	召回率 / %	精确率 / %	F 值 / %
全部特征	79.35	50.42	61.66
-DISTANCE	78.05	52.00	62.42
-I_CONTEXT	78.58	52.13	62.68
-I/J_RELATED_CONTAINS	72.10	56.39	63.29
-I_PRONOUN	74.70	60.95	67.13
-J_DEF	75.64	60.61	67.30
-APPOSITIVE	78.95	59.45	67.82

最先被去掉的是距离特征 DISTANCE,即两个待识别共指对相隔的句子数,说明这个特征在这个语料上表现不理想。

之后是上下文特征 I_CONTEXT。这个特征把先行词 I 周围的词作为放到维基百科中计算。一般先行词很少有代词,而且一般先行词在句首。上下文少了上文,所以结果不太好。相反,指代词 J 是代词的可能性较大,在系统中的效果会好些。所以特征选择过程中没有选择 J_CONTEXT。

接着是语法类特征和维基百科类的 RELATED_CONTAINS。该特征被删除,分析其原因是维基百科的正文里超链接数量众多,使用效果不明显,并最终导致精确率降低。

随后被选择删除的是 I_PRONOUN、J_DEF 和 APPOSITIVE。删除 I_PRONOUN 能大幅度提高结果的精确率,这说明 I_PRONOUN 使得很多共指关系被判断错误,主要原因还是先行词 I 是代词的情况相对较少,从而出现该特征的数据稀疏。J_DEF 和 APPOSITIVE 在最后阶段被选择删除,说明这两个特征相对于前面被删除的特征还是非常有用的。同时,删除这个两个特征,对最终结果的影响很小,仅有 1 %的 F 值。APPOSITIVE 删除后召回率提高精确率降低,也说明 APPOSITIVE 在语料中出现较少,导致数据稀疏。事实上,该特征要求 Mention 对〈I, J〉位置相邻,这种情况肯定所占比例很小。

6　结论与展望

本文通过实验,证明了背景知识对共指消解的增强的作用,结果同时也体现了特征选择算法在机器学习的特征选择过程中发挥的重要作用。从实验结果中可以看出,词汇学特征依旧是目前共指消解中最有效的特征。背景知识的加入,使系统获得了篇章以外的知识,这相当于找到了两个表述之间潜在的联系。在加入背景知识的特征后,系统就能够识别出一些通过普通特征无法确定的共指关系,从而在整体上提高了系统的召回率。

分析实验结果得出结论,利用 WordNet 来计算语义相似度对结果有很大的改善,而对于维基百科的各种特征,还有少部分有没有正面提高。这说明对于维基百科的挖掘还不深入,还有很多地方可以改进。另外,浅层语义关系特征和模板特征都能对结果有一定的提高,这说明语义关系和统计得到的模式对共指消解亦有帮助。在所有背景特征中,上下文特征使精确率提高最大,这说明通过了解表述上下文的词语,能够帮助系统更加精确的识别出共指关系,特别是对于代词这类本身信息量较少的词汇。

(下转第 109 页)

焦点分布的依据。

参考文献：

[1] Teng, Shou-hsin. Remarks on cleft sentences in Chinese[J]. Journal of Chinese Linguistics, 1979, 1: 101-114.

[2] 方梅. 汉语对比焦点的句法表现手段[J]. 中国语文, 1995, 4: 279-287.

[3] 袁毓林. 从焦点理论看句尾‘的’的句法语义功能[J]. 2003, 1, 3-16.

[4] 徐烈炯. 焦点的语音表现[M]. 焦点结构和意义的研究. 北京: 外语教学与研究出版社, 2005, 35-46.

[5] Gussenhoven, Carlos. Focus, mode and the nucleus [J]. Journal of Linguistics, 1983, 19: 377-417.

[6] Gussenhoven, Carlos. Two views of accent: a reply [J]. 1985, 21: 125-38.

[7] 刘探宙. 多重强式焦点共现句式[J]. 中国语文, 2008, 3, 259-269.

[8] Xu, Yi. Effects of tone and focus on the formation and alignment of F_0 contours [J]. Journal of Phonetics, 1999, 27, 55-105.

[9] Chomsky, Noam & Morris Halle. The Sound Pattern of English [M]. New York: Harper and Row. 1968.

[10] Jackendoff, Ray. Semantic Interpretation in Generative Grammar [M]. Cambridge, Mass: MIT Press. 1972.

[11] Selkirk, Elizabeth. Phonology and Syntax: the Relation between Sound and Structure[M]. Cambridge, Mass: MIT Press. 1984.

[12] Rochemont, Michael S. Focus in Generative Grammar [M]. Amsterdam: John Benjamins. 1986.

(上接第 9 页)

另外,最大熵和 SVM 方法在整体实验结果上没有太大差异,只是分别偏重的方面不一致。

本文的方法还有很多地方可以提高,比如提高匹配模式的精度、深入挖掘维基百科以及更加合理的利用上下文信息。另外,还可以继续在该框架下加入新的背景语义知识特征。对共指消解来说,有选择的集成各种深入挖掘的背景语义知识会取得好的效果。

参考文献：

[1] Jun Lang, Bing Qin, Ting Liu, Sheng Li. 2007. Intra-document Coreference Resolution: The state of the art[J]. Journal of Chinese Language and Computing, 17 (4) :227-253

[2] Ponzetto, Simone Paolo and Michael Strube. 2006. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. [C]// Proceedings of the Human Language Technology Conference of the NAACL, Main Conference 2006.

[3] David L. Bean and Ellen Riloff. 2004. Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution. [C]// Proceedings of HLT-NAACL 2004.

[4] Xiaofeng Yang and Jian Su. 2007. Coreference Resolution Using Semantic Relatedness Information From Automatically Discovered Patterns. [C]// Proceedings of ACL 2007.

[5] J. McCarthy and W. Lehnert. 1995. Using decision trees for coreference resolution. In: C. R. Perrault ed. [C]// Proc. of the Fourteenth International Joint Conference on Artificial Intelligence. Qu bec, Canada: Springer, 1050-1055.

[6] Kohavi, R., G. H. John. 1997. Wrappers for feature subset selection[J]. Artificial Intelligence Journal. 97 (1-2): 273 – 324.

[7] 张钹. 2007. 自然语言处理的计算模型[J]. 中文信息学报, 21(3) :3-7.

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

[9] X. Luo. 2005. On coreference resolution performance metrics. [C]// Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 25-32.

阅读此文的还阅读了：

1. 语义信息集成的知识体系探讨
2. 一种面向共指消解的多语义网实体对比表自动生成方法
3. HIPS 6.0集成多种防护机制
4. 一种基于谱聚类的共指消解方法
5. 微波消解——ICP-MS测定多种元素
6. 一类“语义悖论”之消解——作为自我否定句的“语义悖论”
7. 语义识别背后的知识背景
8. 语义级知识融合中的冲突消解方法
9. 一种基于MapReduce的实体共指消解方法
10. 多种语义特征在突发事件新闻中的共指消解研究
11. 消解“悖论”的预设知识--语义、自涉、矛盾、内容
12. 维基百科语义背景知识的共指消解研究
13. 基于语义Web的企业知识集成研究
14. 基于MapReduce的对象共指消解方法
15. 多种学科知识的展演
16. 知识服务与语义检索
17. 共指消解研究方法综述
18. 语义Web中对象共指的消解研究
19. 面向体育新闻领域的中文简单名词短语共指消解
20. 基于语义集成的客户知识挖掘模型研究
21. 消解“悖论”的预设知识——语义、自涉、矛盾、内容
22. 一种基于谱聚类的共指消解方法
23. 基于决策树的汉语代词共指消解
24. 歧义敏感自然语言处理系统中的共指消解
25. 情态模糊语义的消解

26. 基于触发词语义选择的Twitter事件共指消解研究

27. 基于中心语匹配的共指消解

28. 语义韵的相关知识综述

29. 面向产品评论的共指消解方法研究与实现

30. 基于超图分割的共指消解研究

31. 背景知识对语义表达与理解的制约

32. 语义Web作为背景知识的本体匹配

33. 汉语成语语义韵的冲突及其消解

34. 面向知识图谱的共指消解方法研究

35. 浅析消解语义悖论的意义

36. 集成多种背景语义知识的共指消解

37. 格雷林悖论的语义分析及消解

38. 基于待消解项识别的全局优化共指消解方法研究

39. 基于共指消解的实体搜索模型研究

40. 知识的娱乐化与意义的消解

41. 基于语义网的中文百科知识组织与集成

42. 语义技术与知识管理

43. 基于实例动态泛化的共指消解及应用

44. 面向共指消解的动态泛化机制研究

45. 微波消解仪的基础知识和应用前景

46. 词汇语义知识库浅述

47. 语义网络的模糊知识管理

48. 基于最大熵模型的共指消解研究

49. 结合Web背景知识的图像语义标注

50. 基于最大熵模型的共指消解研究