

**编者按：**中国中文信息学会于 2010 年 8 月在哈尔滨工业大学成功地召开了“第六届全国信息检索学术会议(C CIR2010)”。会议的程序委员会向本刊推荐了 23 篇论文,并经作者仔细修改,编辑部得到授权,已在 2011 年第 1 期发表 20 篇,其余 3 篇在本期发表,以飨读者。

文章编号: 1003-0077(2011)03-0003-06

# 基于中心语匹配的共指消解

张牧宇,黎耀炳,秦 兵,刘 挺

(哈尔滨工业大学 计算机学院信息检索研究中心,黑龙江 哈尔滨 150001)

**摘 要：**共指消解是自然语言处理的核心任务之一。在传统机器学习方法使用的平面特征基础上,该文提出一种利用中心语信息的新方法。该方法首先引进一种基于简单平面特征的实例匹配算法用于共指消解。在此基础上,又引入了先行语与照应语的中心语字符串作为新特征,并提出一种竞争模式对将中心语约束融合进实例匹配算法,提升了消解效果。该方法与其他只使用平面特征的传统机器学习方法相比,能充分地利用每一个训练实例的特征信息,进一步融合中心语字符串特征使消解效果更加准确。

**关键词：**中心语匹配;实例匹配;共指消解

**中图分类号：**TP391      **文献标识码：**A

## Coreference Resolution Based on Head Match

ZHANG Muyu, LI Yaobing, QIN Bing, LIU Ting

(Center for Information Retrieval School of Computer Science and  
Technology Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

**Abstract:** Coreference Resolution is one of the core issues in Natural Language Processing. Based on flat features for traditional machine learning method, we propose a new method for exploiting information of the head. Firstly, we introduce an instance-matching algorithm based on simple flat features for coreference resolution. With such instance-matching algorithm, we introduce the head string of antecedent and anaphora as new feature, and propose a competition mode to integrate the head-string feature into instance-matching. Compared to other traditional machine learning methods which just consider flat features, our method can fully exploit the feature information for each training instance and the fusion of head string feature produces more accurate result.

**Key words:** head match;instance match;coreference resolution

## 1 引言

现实世界中同一个事物经常会有不同的名称以及描述,我们称这些名称以及描述称为“表述”(Mention),称这些表述所对应的事物为“实体”(Entity)。所谓的共指消解(Co-reference Resolu-

tion),就是根据一篇文档中各个表述的内容以及上下文信息将这些表述对应到具体实体的过程<sup>[1-2]</sup>。实质上,共指消解是一个对所有表述进行等价类划分的过程,它可以使隐藏在陈述中的等价关系变得清晰,这对于信息抽取、信息检索、机器翻译等上层应用的进一步发展是非常有帮助的。

近 20 年来,共指消解研究受到了特别关注,大

多数计算模型和实现技术都是这一时期出现的。随着 MUC、ACE、ARE 等共指消解相关国际评测相继开展,共指消解得到快速的发展。

对共指消解的研究,无论是早期基于语言学规则的方法,还是近些年来基于统计机器学习的研究体系<sup>[3-4]</sup>,都离不开对高质量特征信息的依赖。但当前 NLP 底层预处理效果不尽如人意,随着添加特征的增多,因前期预处理错误而产生的特征抽取错误也会增多。一些研究者通过特征筛选解决此问题,但这仅限于去除一些准确率较低的特征,仍不能去除那些预处理阶段已经产生的错误对训练数据的污染。

针对这些问题,本文提出基于中心语匹配的共指消解新方法,用于在消解过程中消除预处理过程中产生的部分错误,并捕获一些语义互斥的搭配信息,在使用相同信息的情况下,得到优于传统机器学习方法的共指消解效果。本文的第 2 节介绍共指消解的相关研究工作,第 3 节具体介绍一种简单、有效的基于平面特征的实例匹配算法用于共指消解,第 4 节提出一种“竞争模式”将中心语特征融入到实例匹配算法当中,第 5 节介绍实验设置以及实验结果与分析,第 6 节给出结论。

## 2 相关研究工作概述

共指消解的早期阶段,以基于语言学方法的研究思路为主,代表方法是 Hobbs 算法<sup>[8]</sup>以及中心语理论<sup>[9]</sup>。他们大多使用了大量的语言学规则,而规则的有效性决定于规则的表达能力和可满足性。具体地说,规则前提的可满足性与 NLP 预处理效果密切相关。但即使是目前的研究水平,NLP 底层预处理仍有诸多不如人意的地方。这迫使上世纪末,研究者开始转向各种鲁棒性更强的基于向量相似度计算的聚类方法以及基于二元对的机器学习分类方法。

基于二元对的机器学习分类方法,将共指消解这样一个等价类划分问题转化为任意两个表述是否存在共指关系的问题。在这种典型的二元分类框架<sup>[2]</sup>下,存在共指关系的两个表述中,靠前的都称之为“先行语”,靠后的都称之为“照应语”。

本世纪初以 Soon 等人为代表<sup>[3-4]</sup>,在这种二元分类框架下,采用基于统计的机器学习方法研究共指消解问题。这些方法采用一系列和共指现象相关的信息作为机器学习算法的特征信息,经典的有如 Soon 等人于 2001 年总结提出的 12 个语义、词汇特

征<sup>[3]</sup>。随后 Ng 等人<sup>[4]</sup>于 2002 年对 Soon 使用的 12 个特征进行了扩展,共使用了 53 个特征,涵盖了语义词汇等各个方面,取得了很好的共指消解效果。这一类特征均取离散值或者连续数值,常称之为“平面特征”。

近几年基于机器学习的研究方法开始进一步挖掘利用字符串匹配信息<sup>[5]</sup>,将更多的结构化特征<sup>[6]</sup>融合进机器学习过程,甚至使用更多的背景语义知识作为特征信息<sup>[7]</sup>。有的研究者则尝试使用分类器组合策略<sup>[10]</sup>增强现有特征下的分类器消解效果。

这些根据训练数据特征信息拟合分类模型的机器学习方法,并没有进一步考虑预处理错误对训练数据造成的污染。对这种级联错误的解决办法,NLP 领域目前较流行的解决办法是使用统一的语言模型,但这需要对 NLP 各部分预处理都要较深入的了解,在共指消解这个子任务上尚未有成熟的解决方案。

本文针对以上问题,提出基于中心语匹配的共指消解办法,希望在共指消解的过程中消除预处理过程产生的部分错误,并捕获一些语义互斥的搭配信息,在实验验证中取得了较好的结果。

## 3 基于平面特征的实例匹配算法

为了保证基于中心语匹配的共指消解效果,并且由于本文的研究重点不在于平面特征的丰富性,故本文采用了 Soon 等人总结提出的 12 个平面特征作为基础特征集。

基于平面特征的实例匹配算法,其核心思想是从训练实例中寻找那些与测试实例最相似的实例,并根据相似实例的类别分布给出测试实例属于某个类别的置信度。

本文将实例间的相似度定义为两实例拥有相同特征取值的个数。在寻找那些与测试实例最相似的实例过程,我们希望有不少训练实例的特征取值与测试实例的特征取值完全一致,并且这些实例全为正例(或反例),此时便能以置信度 1 将测试实例判定为正例(或反例)。但实际情况中常常出现两类偏差:

(1) 不存在特征取值与测试实例完全一致的训练实例。

对于这类偏差,可称之为匹配失效,产生的原因有二:其一,测试实例某个特征取值未曾在训练实例中出现过;其二,测试实例的特征取值在不同的训

练实例上出现过,但没有任何一个训练实例的特征取值与测试实例完全一致。总的说来,这是由于数据稀疏造成的。

(2) 特征取值与测试实例完全一致的训练实例的类别分布过于均匀。

对于这类偏差,一方面,在于现有特征的表现力不足,不足以有效区分两种类别;另一方面,在于训练数据受到污染,特征抽取不准确。

由于以上两类偏差的存在,使用全部特征取值挑选与测试实例最相似的训练实例是不明智的,需要对完全匹配的要求进行修改以适应实际应用。本文设计的基于平面特征的实例匹配算法,利用决策树算法<sup>[11]</sup>中使用的“信息熵”(Information Entropy)作为选择函数,通过贪婪思想挑选与测试实例最相似的训练实例。

“信息熵”描述了一个带类别标签的样本集中类别分布的均匀程度,“信息熵”越大表明各类别样本的数量差别越小,从中随机抽取一个样本并成功将其判定为某一类的概率越低。在样本集  $S$  中只有两类情况下,若样本集中类别为“ $P$ ”(Positive)和“ $N$ ”(Negative),相应的样本数量分别为  $p$  与  $n$ ,则“信息熵”具体表示如下:

$$I(S) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (1)$$

可以证明,当  $p=n$  时,信息熵为 1;而  $p=0$  或  $n=0$  时,信息熵为 0。

以下具体描述基于平面特征的实例匹配算法:

#### 算法 1 基于平面特征的实例匹配算法

输入: 抽取了平面特征的训练实例库、待测试实例

输出: 待测试实例的正例置信度

- (1) 定义“特征取值对”为“<特征名, 取值>”, 测试实例  $i$  的所有“特征取值对”构成未使用的特征取值列表  $L$ , 所有训练实例构成候选实例集  $D$ ;
- (2) 从  $L$  中选择一个“特征取值对” $e$ , 使得实例集  $D$  中满足该  $e$  要求的子集  $D'$  的信息熵  $I(D')$  最小, 并令  $D=D'$ , 从  $L$  中删除  $e$ ;
- (3) 若  $I(D)=0$  或  $L$  为空, 返回训练实例子集  $D$ , 及其中正例所占比例  $p/(p+n)$  作为测试实例  $i$  的正例置信度; 否则返回步骤(2)。

可以看到以上基于平面特征的实例匹配算法将寻找匹配实例的过程, 转化为通过贪婪方式选取“特征取值对”并筛选训练实例的过程, 最终算法返回了局部优化的匹配结果, 并通过训练实例的类别分布给出正例置信度。

该实例匹配算法设计思路简单, 但如表 2 和

表 3 所示的实验结果所示, 在简单平面特征集上取得了与传统机器学习方法 Maxent(最大熵)、J48(决策树)和 SVM(支撑向量机)相当的共指消解效果。

对实验结果进行详细分析后发现, 上述算法 1 产生的错误中, 存在命名实体识别错误和先行语与照应语语义互斥两类现象:

#### (1) 命名实体识别错误

对于同一领域的语料或者文本, 常使用同一命名实体识别(NER)工具及名词短语抽取工具从文本中抽取待消解的候选表述。但由于这些工具往往出现不同程度的识别错误, 而相同的识别错误会反复发生在训练和测试阶段。如我们在实验时曾发现, 现使用的 NER 工具及名词短语抽取工具常将 ACE2005 英文语料上的“news”识别为候选先行语或照应语, 但“news”单词不在 ACE 语料考察的实体类型范围之内。类似的错误还包括曾反复将“the game”、“tuesday night”等作为照应语。

#### (2) 语义互斥的搭配

在 ACE 新闻语料中, “We”与“them”分别是第一和第三人称代词, 而前者常指新闻报道员或评论员, 绝大多数情况下两个表述不应该存在共指关系。当这两个表述在同一句子出现, 且明显满足单复数一致性, 这导致在缺少更多背景知识的情况下, 无法正确将其判定为反例, 进而做出错误的决定。

事实上, 这两类错误信息可以通过后续融合中心语特征的实例匹配算法加以捕获利用。

## 4 融合中心语特征的实例匹配算法

如上节所述, 基于平面特征的实例匹配算法无法有效识别“命名实体训练错误”和“语义互斥的搭配”两类错误。通过增强命名实体与名词短语抽取模块可以减少第一类错误, 但短期内这类错误很难得到彻底解决; 对于第二类错误, 容易想到的解决办法是从训练数据中挖掘这些语义互斥的搭配模板, 但模板挖掘依赖于频率计算, 一些出现次数较少但语义上同样明显不相容的名词短语对很难被充分挖掘。

实际上, 通过实例匹配算法直接查找那些与测试实例拥有相同中心语的训练实例, 则能根据这些训练实例的类别分布做出正确的判断, 从而有效地解决这两类问题。下面进一步通过强制约束与协商约束两种方式, 将中心语约束信息融合进实例匹配算法。

4.1 基于中心语的强制约束

如上所述,通过中心语字符串的表面约束能正确识别出部分反例,但实际上融合中心语特征与平面特征还能增加一些隐含的匹配约束信息,这是实例匹配过程所需要的。有鉴于此,我们在此增加两个中心语字符串特征,具体分别为“AntHead”(先行语的中心语字符串)和“AnaHead”(照应语的中心语字符串),其直接以字符串的形式作为一个特征取值,具体内容和长度视情况而定,对长度没有限制。如上面的实例“<We,them>”,其对应的中心语特征分别为“AntHead = We”和“AnaHead = them”。

基于中心语强制约束的实例匹配算法如下:

算法 2 基于中心语强制约束的实例匹配算法

输入:抽取了平面特征和中心语特征的训练实例库、待测试实例

输出:待测试实例的正例置信度

- (1) 定义“特征取值对”为“<特征名,取值>”,测试实例  $i$  的所有“特征取值对”(不包括中心语特征)构成未使用的特征取值列表  $L$ ,所有训练实例构成候选实例集  $D$ ;
- (2) 从  $D$  中选择与测试实例在“AntHead”特征上取相同值的所有实例子集  $D'$ ,如果  $D'$  为空集,直接跳到(4);否则,令  $D=D'$ ;
- (3) 从  $D$  中选择与测试实例在“AnaHead”特征上取相同值的所有实例子集  $D'$ ,如果  $D'$  为空集,直接跳到(4);否则,令  $D=D'$ ;
- (4) 从  $L$  中选择一个“特征取值对” $e$ ,使得实例集  $D$  中满足  $e$  要求的子集  $D'$  的信息熵  $I(D')$  最小,并令  $D=D'$ ,从  $L$  中删除  $e$ ;
- (5) 若  $I(D)=0$  或  $L$  为空,返回训练实例子集  $D$ ,及其中正例所占比例  $p/(p+n)$  作为测试实例  $i$  的正例置信度;否则返回步骤(4)。

与算法 1 相比区别在于,算法 2 首先利用“Ant-Head”和“AnaHead”对训练实例集进行筛选,随后再使用平面特征进一步挑选实例。为了验证算法 2 的效果,我们在小数据集(ACE2005 英文 BNews 语料,16 篇训练语料,10 篇测试语料)上测试了的共指消解效果,具体结果见表 1。

表 1 小数据集上的实验结果

特征集	MUC-6 评测指标		
	R/%	P/%	F/%
Basic	61.1	52.0	56.2
Basic+Head	68.1	30.8	42.4

如表 1 所示,“Basic”表示 Soon 等人提出的 12

个平面特征,“Basic+Head”表示 12 个特征再加上前面定义的“AntHead”、“AnaHead”两个中心语特征。可见,虽然增加了中心语特征,算法 2 在 MUC-6 评测标准的  $F$  值上比算法 1 低很多。显然,这种组合使用平面特征与中心语特征的方式过于粗糙,下面我们将针对出现的一些问题对其进行改进。

4.2 基于中心语的竞争模式

4.2.1 强制约束的错误分析

对算法 2 与算法 1 的结果进行详细对比后发现:虽然算法 2 能够有效解决以上所述两类错误,但由于算法 2 中步骤(2)、(3)硬性引入中心语过滤操作,反而使噪声信息被放大。具体地说,算法 2 引入中心语过滤操作后,产生两类差异:

(1) 坏的差异:原被算法 1 正确分类的实例,现被算法 2 错误分类。

这主要是由于中心语约束过于严格,经过中心语过滤后剩余的训练实例偏少,呈现严重的数据稀疏性。后续以信息熵为选择函数通过贪婪方式选取“特征取值对”,会使过滤后剩余的实例变得更加少,并且此时类别分布极不均匀。这相当于只使用很少的平面特征便给出类别判定,一旦这几个实例的特征取值出现噪声,便造成类别判定出现截然相反的结果。

(2) 好的差异:原被算法 1 错误分类的实例,现被算法 2 正确分类。

这主要是由于算法 2 专注于解决上述两类错误,并希望引入更多隐含的深层匹配约束信息所得到的收益,是我们所希望看到的。

4.2.2 竞争模式

我们希望去除大部分“坏的差异”,保留“好的差异”。在一定的条件下我们使用算法 2 的结果,而在另外的情况下我们使用算法 1 的结果,以这种思想为基础,我们引入“竞争模式”这种相对柔和的协商约束方式。

算法 3 竞争模式

- (1) 在 Basic+Head 特征集上利用算法 2 获取与测试实例最匹配的训练实例子集  $D_2$ ;
- (2) 在 Basic 特征集上利用算法 1 获取与测试实例最匹配的训练实例子集  $D_1$ ;
- (3) 计算训练实例子集  $D_1$  与  $D_2$  的“重要度”, $ScoreIMP(D_1)$  和  $ScoreIMP(D_2)$ ;
- (4) 如果  $ScoreIMP(D_1) > ScoreIMP(D_2)$ ,返回  $D_1$  中正例所占比例作为测试实例的置信度;否则,返

回  $D_2$  中正例所占比例作为测试实例的置信度。

重要度计算公式：

$$ScoreIMP(D) = \left( \frac{2}{1 + e^{-\alpha \times n_D}} - 1 \right) \times |r_D - 0.5| \tag{2}$$

其中：(1)  $n_D$  为训练实例子集  $D$  中全部实例的数量；  
(2)  $r_D$  为训练实例子集  $D$  中正例所占比例。

注意到公式(2)中  $\alpha$  为大于 0 的待定参数，其值取决于训练数据的具体分布，在实验中采用交叉验证获得其最优值。从公式(2)可见，当  $D$  中实例数量越多，且正例所占比例越接近 0 或 1 时(此时类别分布非常不均匀)，评分越高。它是综合考虑“信息量”(支撑结论的证据数)和“类别分布均匀度”的结果。

上述竞争模式就是我们将中心语特征融合进实例匹配算法的方式。实验表明，竞争模式算法作用在“Basic+Head”特征集上，能有效去除绝大部分“坏的差异”，保留为数不少的“好的差异”，从而增强共指消解的效果。

### 5 实验结果以及分析

我们通过实验验证了前述“基于平面特征的实例匹配算法”与“融合中心语特征的实例匹配算法”的效果，下面介绍实验设置与具体的实验结果。

#### 5.1 实验设置

本文在 BART 共指消解平台上完成实验<sup>[12]</sup>，该平台首先对语料进行预处理，包括 Tokenization、NER、POS、Parsing 等。

我们使用 ACE2005 英文 BN 和 NW 语料作为实验数据，其分别包含 226、106 篇语料。使用的特征包括 Basic 特征集与 Head 特征集，具体见 4.1 节中的描述。

训练实例的构造方式与测试过程中使用的消解策略均采用 Soon 等人<sup>[3]</sup>提出的方法，并采用 MUC-6 评价方法中  $R$ 、 $P$ 、 $F$  三个指标表征消解效果。由于公式(2)中存在待定参数，故采用 5-fold 交叉验证方式验证不同参数下的共指消解效果。

#### 5.2 实验结果

在基本特征集(Basic)及扩展特征集(Basic+Head)下，表 2 和表 3 分别给出了利用 ACE2005 英

表 2 ACE2005 BN 语料上实验结果

二元分类方法	特征集	R/%	P/%	F/%
Maxent	Basic	53.5	50.9	52.1
J48	Basic	52.5	51.6	52.0
SVM	Basic	50.2	50.7	50.5
A1	Basic	56.7	48.6	52.3
A3	Basic+Head	54.7	53.2	53.9

表 3 ACE2005 NW 语料上实验结果

二元分类方法	特征集	R/%	P/%	F/%
Maxent	Basic	53.2	61.7	57.1
J48	Basic	53.3	63.0	57.7
SVM	Basic	52.7	62.4	57.1
A1	Basic	53.7	62.4	57.7
A3	Basic+Head	53.3	71.6	61.1

文 BN 与 NW 语料的实验结果，其中 A1 和 A3 分别表示算法 1 和算法 3，而 A3 在表 2 和表 3 中的结果都在最优  $\alpha$  值下取得的结果( $\alpha$  最优值分别为 0.6 和 0.5)。

结果表明，在基本特征集下，无论是使用 BN 语料还是 NW 语料，基于平面特征的实例匹配算法(A1)在  $F$  值上都达到了 Maxent、J48 和 SVM 三种传统机器学习算法的最好效果。这证实了 A1 算法虽然简单但有效，通过寻找与测试实例最相似的训练实例，能够使共指消解效果达到传统机器学习模型的最好水平。

进一步观察表 2 和表 3，发现融合中心语特征的实例匹配算法(A3)与传统的 Maxent、J48 和 SVM 算法相比，均能明显提高  $F$  值。在 BN 语料上  $F$  值提高幅度达 1.8%~3.4%，而在 NW 语料上  $F$  值提高幅度达 3.4%~4.0%。并且可以观察到，A3 在  $F$  值上的提高很大程度是由于  $P$  值得到的较大的提升，提升幅度分别为 1.6%~2.5%、8.6%~9.9%。算法 3 提出的竞争模式的初衷是将中心语特征有效地融合进实例匹配算法，使其能有效捕获利用 4.1 节中提出的两类信息，进而减少  $P$  值的损失。实验结果可以证实，算法 3 达到了我们预期的效果，使  $P$  值得以提升，融合中心语特征的实例匹配算法通过使用中心语约束信息增强了共指消解效果。

## 6 结论

本文通过实验,证明了中心语匹配约束对共指消解的增强作用。结果同时也体现了基于简单平面特征的实例匹配算法简单而有效,能充分利用每一个训练实例的特征信息,其效果达到了传统机器学习方法的最优水平。进一步,实验结果证明,引入中心语字符串特征并使用竞争模式后,基于中心语匹配确实能有效去除预处理过程中产生的部分错误及捕获一些语义互斥的搭配,提高了共指消解的 P 值,从而改善的共指消解的总体效果。

本文的方法还有很多地方可以提高,比如算法 1 通过贪婪方式选择“特征取值对”过滤训练实例只能取得局部最优的选择序列,要获得更好的选择序列应该从全局信息出发修改算法。另外,考虑照应语和先行语的修改语作为字符串特征,增加约束信息会取得更好的效果。

## 参考文献

- [1] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报, 2002, 16(6):9-17.
- [2] J. Lang, B. Qin, T. Liu, et al. Intra-document Coreference Resolution: The state of the art[J]. Journal of Chinese Language and Computing, 2007, 17(4): 227-253.
- [3] W. M. Soon, H. T. Ng, D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521-544.
- [4] V. Ng, C. Cardie. Improving Machine Learning Approaches to Coreference Resolution[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics, 2002:104-111.
- [5] X. Yang, G. Zhou, J. Su, et al. Improving Noun Phrase Coreference Resolution by Matching Strings [C]//Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP2004), Hainan Island, China, 2004. 2004:22-31.
- [6] X. Yang, J. Su, C. L. Tan. Kernel-based pronoun resolution with structured syntactic knowledge[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006. Association for Computational Linguistics, 2006:41-48.
- [7] 郎君, 析舟, 秦兵, 等. 集成多种背景语义知识的共指消解[J]. 中文信息学报, 2009, 23(3):3-10.
- [8] J. R. Hobbs. Resolving Pronoun References[J]. Lingua, 1978, 44:311-338.
- [9] B. J. Grosz, S. Weinstein, A. K. Joshi. Centering: A framework for modeling the local coherence of discourse[J]. Computational Linguistics, 1995, 21(2): 203-225.
- [10] S. Vemulapalli, X. Luo, J. F. Pitrelli, et al. Classifier Combination Techniques Applied to Coreference Resolution [C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium, Boulder, Colorado, 2009. Association for Computational Linguistics, 2009:1-6.
- [11] J. Quinlan. Induction of decision trees[J]. Machine learning, 1986, 1(1):81-106.
- [12] Y. Versley, S. P. Ponzetto, M. Poesio, et al. Bart: A modular toolkit for coreference resolution [C]//Proceedings of the ACL-08: HLT Demo Session, Columbus, Ohio, 2008. Association for Computational Linguistics, 2008:9-12.