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)03-0010-07

一种基于谱聚类的共指消解方法

谢永康,周雅倩,黄萱菁

(复旦大学 计算机科学与工程系,上海 200433)

摘 要: 该文针对中文共指消解的具体任务,提出采用谱聚类的方法进行共指消解。首先,在待消解项对上抽取特征,使用最大熵模型判断两个待消解项存在共指关系的概率;然后,以此概率值作为相似度进行谱聚类;最后,得到若干实体,实现共指消解。该方法能从全局的角度进行实体划分,有效地提高准确率。在 ACE 2007 标准数据集上的 Diagnostic 实验结果表明该方法的 ACE Value 比 baseline 方法有了 2.5%的提高,Unweighted Precision 值有 5.4%的提高。

关键词: 计算机应用;中文信息处理;共指消解;谱聚类;最大熵模型

中图分类号: TP391 **文献标识码**: A

A Spectral Clustering Based Coreference Resolution Method

XIE Yongkang, ZHOU Yaqian, HUANG Xuanjing

(Department of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: This paper presents a novel method to implement coreference resolution. This method is based on spectral clustering. A maximum entropy model is first used to get the coreference probability of mention pairs with extracted features. The probabilities of mention pairs are then used to construct the similarity matrix for spectral clustering. Entities are generated according to the clustering cuts. This method can divide entities with a global view, which effectively improves precision. Experiments on ACE 2007 dataset show that the ACE Value of this method is 2.5% higher than that of baseline on Diagnostic task, and 5.4% higher in Unweighted Precision.

Key words: computer application; Chinese information processing; coreference resolution; spectral clustering; maximum entropy model

1 引言

近年来,越来越多的学者开始关注指代消解的研究[1]。指代一般分为回指(Anaphora)和共指(Coreference)。共指主要是指多个命名实体,名词短语,名词或代词指向现实世界中的同一实体。共指消解的目标是识别出文档中所有存在的共指关系。由于共指消解在问答系统、信息检索和机器翻译等自然语言处理任务中有广泛的应用,因此,共指消解成为了研究的热点[2],特别是在中文的共指消解的研究中,越来越多的专家学者提出了不同的消解算法[3-5]。这些算法总体上可分为有监督学习和无监督学习两大类。

目前,较多的共指消解算法属于有监督学习这一类。这类方法通常先使用统计学习的分类器判断待消解项对的共指关系,然后,采用不同的共指链归并或聚类算法实现共指消解。下面列举共指消解中的一些经典聚类算法,并给出一些分析。

Soon et. al.[6]和 Ng and Claire[7]对每个待消解项采用从右到左寻找可能的先行词的方法将待消解项归并成若干个实体。不同的是,文献[6]中的方法

收稿日期: 2008-08-28 **定稿日期**: 2008-11-02

基金项目: 国家自然科学基金资助项目(60503070);技术发展高校资助项目(GH0742002)

作者简介: 谢永康(1982—),男,研究生,主要研究方向为自然语言处理;周雅倩(1976—),女,讲师,主要研究方向为自然语言处理;黄萱菁(1972—),女,教授,主要研究方向为自然语言处理。

使用从右到左第一个可能的先行词,即 pick-first[8] 的算法;而文献[7]中的方法使用从右到左的先行词中可能性最大的那个进行归并,即 best-first[8] 的算法。

虽然,上述两个方法的计算复杂度低,但对于分类器的判断错误比较敏感。假设 A, B, C 这三个待消解项从左到右排列。如果分类器的输出结果中将 A 和 C 以及 B 和 C 判断为共指,而 A 和 B 没有被判断为共指,那么,对于待消解项 C,它或与 A 共指或与 B 共指,但无法构成包含 A,B,C 三个待消解项的共指链。对于图 1 中的例子,文献[6-7]中的方法会将所有的待消解项归并成一个实体。图 1 中的数值代表两个待消解项存在共指关系的概率。

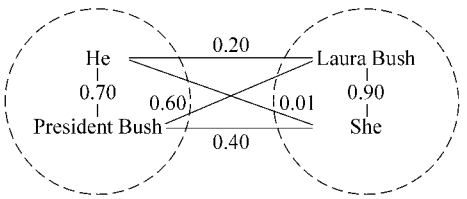

图 1　共指消解示例

Florian et al.[8] 和 Luo et al.[9] 提出使用 Bell Tree 的聚类算法。它们将从左到右每一个待消解项对应到 Bell Tree 上的从根到叶子的一层,每层中的节点代表一种包含当前待消解项及其左面所有待消解项构成的实体划分情况。通过计算树的根节点到叶节点的分值,该算法决定采取归并待消解项、产生新的实体或对 Bell Tree 剪枝的动作。最终找到从根到叶节点的最优路径,即实体的最优划分。从而完成共指消解。

该方法能从全局的角度出发搜索最优的实体划分,但搜索空间非常大。虽然,文献[8-9]中的方法提出了相应的剪枝办法以缩小搜索空间,但是,计算复杂度仍然较高。而且对于图 1 中的例子,也会将所有的待消解项划为一个实体。

文献[2]在对待消解项进行聚类时首先从右向左找到当前待消解项的最可能的先行词,并且要求该可能性的值大于阈值 MaxPro,然后,检测当前待消解项与包含先行词的共指链中的任意一个指代项共指可能性是否大于阈值 MinPro。仅在上述两个条件都满足的情况下,才将当前的待消解项与已存在的共指链归并。该方法也能正确划分图 1 中的例子。

文献[5]在进行指代项的归并时使用的是计算待消解项的传递关系闭包的方法。该方法在一定程度上能够避免文献[2,6-7]中归并共指链所遇到的问题。但是存在这样的问题,假设 A, B, C 这三个待消解项是从左到右排列的,且 A 实际上属于一个实体,而 B 和 C 属于另一个实体。那么如果 A 和 C 被判为共指,B 和 C 也被判为共指,那么经过传递闭包 A, B 和 C 将错误地共指同一个实体。

Aron Culotta[10] 提出基于词集(Noun Phrase Cluster)的一阶共指消解概率模型。该方法最终通过层次聚类的方法由待消解项得到实体。通过 B-Cubed 评测表明,相对仅在待消解项对上取特征的方法,该方法能有效地提高召回率(Recall)的值。文献[10]中的方法能够正确地解决图 1 中的例子,但是它的算法复杂度较高。

文献[4]所采用的图切分的聚类算法的相似度计算函数本质是基于规则的,其有效性和准确性比较难以度量。

本文所使用的方法属于有监督学习方法。首先,训练一个最大熵二值分类器,然后,基于分类器输出的待消解项对存在共指关系的概率值进行谱聚类,最终,将文档中存在的共指关系识别出来。不同于文献[6-7]中采用的局部的归并方法,谱聚类能从全局的角度对实体进行划分,通过阈值来控制不同类型的实体所包含指代项的平均个数。该方法能正确地解决图 1 中的例子,同时,该方法的搜索空间和计算复杂度也分别较文献[8-10]中的方法低。通过在 ACE 2007 语料上的实验表明,该方法的 ACE Value 值比 Baseline 方法有了 2.5 %的提高。

第 2 部分详细介绍本文采用的方法,第 3 部分是实验,最后是总结。

2　共指消解

本文提出的基于谱聚类的共指消解算法,是将最大熵分类器输出的待消解项对存在共指关系的概率值作为相似度来构建待消解项的邻接矩阵。采用概率值是为了将分类器学习所得信息加入到最后的指代消解中,以克服通过规则方法得到相似度的局限性。然后,使用 2-way 目标函数 Ncut 将图递归地切成若干个子图,直到子图满足一定条件为止。在该方法中,两个待消解项之间的相似度是通过在训练语料上使用最大熵模型学习得到的。当然,我们也可以使用其他有监督的统计学习方法得到类似的相似度值。通过第一阶段的待消解项识别结果或基于正确的待消解项,我们可以得到每个待消解项的类型。我们仅对属于同种类型的待消解项构建各自

的邻接矩阵,从而有效地降低了算法的运行时间。

2.1 特征选取

通过待消解项识别和分类等预处理后,我们把待消解项按照它们的类别构成两两的待消解项对。针对待消解项对,我们抽取一系列的特征,并用最大熵模型进行训练。

正如文献[3]中提到的,特征的选取对于最大熵分类模型有重要影响。本文将所提取的特征分为 6 大类,分别列在表 1 中。其中,编辑距离用来避免共指待消解项可能存在的拼写错误。字符串匹配类型对于缩写和简称的待消解项有较好的指示作用。在距离特征中,我们选取了两个待消解项之间的距离和两个待消解项所在的句子之间的距离这两个特征。如果两个待消解项 A 和 B 之间存在的词或标点少于两个,且这两个词一个为人名,另一个为名词类型的人称,那么 A 和 B 这两个待消解项很可能存在共指关系。例如,"学习委员汤伟杰"中,"委员"和"汤伟杰"之间不存在其他词,且它们一个为名词类型的人称,另一个为人名,所以很可能存在共指关系。另外,对于两个所在句子之间的距离超过三句的待消解项或者不太可能存在共指关系或者它们之间有另一个待消解项和它们多存在共指关系。

表 1 待消解项对抽取特征

特征类别	特征名称	特征描述
词特征	WL	两个待消解项中左面的词
	WR	两个待消解项中右面的词
词性特征	PosL	左面待消解项中每个词的词性
	PosR	右面待消解项中每个词的词性
词法特征	ED	两个待消解项之间的编辑距离是否小于 2
	SubStrType	两个待消解项之间的字符串匹配类型,包括完全匹配、左子串、右子串、包含、分散存在和不匹配这六种
距离特征	TD	两个待消解项之间存在词的个数是否小于 2
	SenWin	两个待消解项是否在位于连续的三个句子以内
待消解项特征	MTL	左面待消解项的类型,包括 NAM、NOM 和 PRO 分别表示命名实体类型待消解项、名词类型待消解项和代词类型待消解项。
	MTR	右面待消解项的类型。
其他特征	IsAlias	待消解项之间是否存在地名的别称关系
	IsIndividual	两个待消解项是否可能共指同一个人名。如果待消解项共指的实体类型为"PER_Individual",且左面待消解项是名词类型,右面待消解项是命名实体类型,TD 特征为 True,则 IsIndividual 为 True,否则为 False。
	IsGroup	两个待消解项是否可能共指同一个团体。如果待消解项共指的实体类型为"PER_Group",且待消解项的词相同或其中有一个待消解项的词性为"ng",则 IsGroup 为 True,否则为 False。
	IsGPE	两个待消解项是否可能共指同一个地域政治实体。如果待消解项共指的实体类型为"GPE_Nation"或"GPE_Special",且 IsAlias 特征为 True,则 IsGPE 为 True,否则为 False。

待消解项特征是通过待消解项的识别和分类得到的。例如,如果两个命名实体类型(NAM)的待消解项它们的词完全相同,那么它们很有可能存在共指关系。

在其他特征中,我们收集了一些常见的地名的别称列表,例如,"美利坚合众国"和"美国","上海"和"沪"等。该特征在文献[3]中也被提到。通常情况下,人名,地名和团体一般是出现频率最高的,因此,我们特别添加了三个与此有关的特征。

特征模板是由表 1 中的特征组合产生的。例如,特征模板 MTL_MTR_TD 代表通过查看当前待消解项对的 MTL、MTR 和 TD 来决定这两个待消解项是否共指。我们为分类器设计了 14 个由表 1 中的特征组合而来的特征模板。

2.2 最大熵模型

根据待消解项对的特征,我们使用最大熵模型对其进行二值分类,即使用最大熵模型来判断两个待消解项是否存在共指关系,并且输出存在共指关系的概率值,这个值是介于 0 和 1 之间的一

个实数。

自然语言处理等许多领域,最大熵模型[11]已得到了成功的应用。该模型遵循"对已知的建模,对未知的不做任何假设"的原则,使得到的统计模型能满足所有已知的事实,并对未知的事实不做任何假设。该模型的形式是:

$$p(y \mid x) = \frac{1}{Z(x)} \exp \sum_i \lambda_i f_i(x, y) \quad ,$$

其中

$$Z(x) = \sum_y \exp \sum_i \lambda_i f_i(x, y) \qquad (1)$$

这里,x 可以看作待消解项对的特征,即按照表 1 中组合出的特征模版取值后的结果,$p(y \mid x)$ 可看作是给定某个词对,存在或不存在共指关系的概率。$f_i(x, y)$ 是最大熵模型中的特征函数,λ_i 是特征函数的权重,代表每个特征函数的重要性,$Z(x)$ 是归一化因子。

本文中采用 L-BFGS 算法进行最大熵的参数估计。

2.3　谱聚类

谱聚类方法的理论基础是谱图理论[12-13],它以图的 Laplacian 矩阵作为分析基础,利用矩阵和线性代数来研究图的邻接矩阵,根据矩阵的谱确定矩阵的一些性质。

设一无向加权图 $G = \langle V, E \rangle$ 由顶点集合 V 和边的权值集合 E 构成。将 G 表示为对应的对称矩阵 $W[w_{ij}]_{n \times n}$,w_{ij} 表示连接顶点 i 和 j 的边的权值。该图对应的 Laplacian 矩阵为

$$L = D - W \qquad (2)$$

其中,D 为对角阵,$d_{ii} = \sum_j w_{ij}$。

因为,Laplacian 矩阵是对称半正定矩阵,所以,它的所有特征值为非负实数。

根据"类内距离最小,类间距离最大"的原则,Shi 和 Malik[13]提出了将图划分为两个子图的 2-way 目标函数 $Ncut$:

$$\min Ncut(A, B) = \frac{cut(A, B)}{vol(A)} + \frac{cut(A, B)}{vol(B)} \qquad (3)$$

$$cut(A, B) = \sum_{i \in A, j \in B} W_{ij} \qquad (4)$$

$$vol(A) = \sum_{i \in A} \sum_j W_{ij} \qquad (5)$$

其中,$cut(A, B)$ 是切分后子图 A 和 B 间的边,又称为"切边集"。

令 P_j 为 A 的划分指示向量:

$$P_j = \begin{cases} 1, & j \in A \\ -1, & j \in B \end{cases}, \text{其中 } B = A^c \qquad (6)$$

结合公式(2)我们可以得到

$$cut(A, B) = f(p) = \frac{1}{2} p^T L p \qquad (7)$$

考虑约束 $x^T We = x^T De = 0$,则可将公式(3)转化为

$$\min Nuct(A, B) = \min \frac{x^T (D - W) x}{x^T D x} \qquad (8)$$

将 x 松弛到连续域[-1,1],则公式(8)可写为

$$\arg \min_x \min_{x^T De = 0} \frac{x^T (D - W) x}{x^T D x} \qquad (9)$$

根据 Rayleigh 商原理,公式(9)的优化问题等于下式中第二最小特征值的求解问题:

$$D^{-\frac{1}{2}} (D - W) D^{-\frac{1}{2}} x = \lambda x \qquad (10)$$

利用公式(10)中第二最小特征值对应的特征向量 x_2 结合规则划分无向加权图 G。

谱聚类方法最早用于图像的分割。近年来,谱聚类的方法也被用于生物信息学和自然语言处理中。据我们所知,目前还没有在共指消解中应用谱聚类方法的文献。在谱聚类的过程中,我们采用了上述的 2-way 目标函数 $Ncut$。由于这个目标函数体现了"类内距离最小,类间距离最大"的原则,因此,在共指消解的过程中,我们的方法能从全局的角度对由待消解项构成的无向加权图做出最优的划分。我们具体的算法描述如下:

1. 以最大熵模型对两两待消解项是否存在共指关系的概率,作为两个待消解项之间的相似度,构建相似度矩阵 W_i,该矩阵是一个对称阵。其中,i 是实体子类类型的下标。根据 ACE 2007 的定义,共有 45 种子类。我们为一篇文档中属于同一个实体子类类型的待消解项构建一个相似度矩阵。

2. 对每一个矩阵 W_i,将其中所有的待消解项看作一个整体,根据该子类类型的谱聚类阈值,对矩阵 W_i 进行 2-way 的 $Ncut$,从而形成若干个实体。每次划分时我们使用这样的规则:首先根据切边集来判断某个待消解项应该属于哪个实体,如果待消解项位于切边集中,我们将该待消解项归到当前包含指代项个数较少的那个实体中。

3. 将所有的 W_i 中的实体输出。

3　实验

实验采用 ACE 2007 中文评测语料。语料分为两部分,其中 633 篇是训练语料,另外有 256 篇测试语料。分别包含三种来源:广播(broadcast news,

bn, 298 篇)、新闻(newswire, nw, 238 篇)和博客(weblog, wl, 197 篇)。每个来源都包含七大类的实体,分别为 FAC(Facility)、GPE(Geo-Political-Entity)、LOC(Location)、ORG(Organization)、PER(Person)、VEH(Vehicle)和 WEA(Weapon),每个大类又分为若干子类,子类共有 45 个。我们用"大类_子类"的方式来表示实体所属的子类型。所以,一个待消解项所共指的实体类型属于 45 种子类之一。

我们的实验分为两类,一类是在测试语料的待消解项正确识别和分类(即待消解项识别和分类的标准答案)的基础上进行共指消解,即 Diagnostic 评测。另一类是先在测试语料上进行待消解项识别和分类,然后基于这个系统识别结果进行共指消解。我们将后者称为 Open Test 结果。这里的待消解项识别和分类的工具是我们自行设计的。

在实验结果的比较中,我们的 Baseline 和谱聚类方法所使用的分类模型和相关特征是一致的,区别在于最后归并共指链的方法。所有实验数据都是通过 ACE 2007 提供的标准评测工具得到的。从实验效果来看,谱聚类的方法有效地提高了实体划分的准确率。

3.1 Baseline 结果

实验采用的 Baseline 所使用的方法是先通过最大熵分类模型判断两个待消解项是否存在共指关系,然后通过计算待消解项之间的传递关系闭包的方法得到若干共指链。

Baseline 的实验结果列在表 2、3、4 和 5 中,表 2 显示了 Diagnostic 评测中每种语料来源上的结果。由于训练语料本身的缘故,bn 类型的训练比较充分,所以成绩最好。表 3 显示了 Diagnostic 中按照每种实体的大类上取得的结果。从表中的数据中可以看出 Baseline 的召回率(Recall)相对较高,但是准确率(Precision)相对低一些。这是由于最大熵的分类结果中,被判断成具有共指关系的待消解项较少。ACE Value 是对在不同的待消解项类型加上各自的权重(NAM 为 1.0、NOM 为 0.5、PRO 为 0.1)后得到的统计值。从 ACE Value 的总体结果来看,我们的 Baseline 比 ACE 2007 评测时 EDRDiagnostic 任务的最好成绩高 2.2%。这是因为最大熵分类器的判断出的共指关系的准确率比较高。

表 2 Baseline(Diag)按语料结果

语料	Unweighted			ACE
类型	Pre	Rec	F	Value
bn	85.9%	90.5%	88.1%	76.5%
nw	85.5%	89.0%	87.2%	74.5%
wl	81.5%	88.1%	84.7%	68.7%
总体	84.5%	89.1%	86.8%	73.7%

表 3 Baseline(Diag)按实体结果

实体	Unweighted			ACE
类型	Pre	Rec	F	Value
FAC	86.9%	93.5%	90.1%	79.9%
GPE	86.1%	90.9%	88.4%	79.2%
LOC	91.2%	95.7%	93.4%	86.2%
ORG	85.0%	92.4%	88.6%	79.0%
PER	82.4%	85.0%	83.7%	63.6%
VEH	74.7%	88.9%	81.2%	49.6%
WEA	92.7%	89.4%	91.0%	79.9%
总体	84.5%	89.1%	86.8%	73.7%

在 Open Test 上的结果如表 4 和表 5 所示。我们的待消解项的识别和分类结果的 ACE Value 评测成绩为 78.3%,这个成绩在 ACE 2007 的 EMD 评测中排在第四位。比第三位的 82.1%低 3.8%,

表 4 Baseline(Open Test)按语料结果

语料	Unweighted			ACE
类型	Pre	Rec	F	Value
bn	53.9%	58.2%	56.0%	52.6%
nw	53.6%	58.0%	55.7%	49.8%
wl	44.0%	47.4%	45.6%	43.1%
总体	51.2%	55.3%	53.2%	49.0%

表 5 Baseline(Open Test)按实体结果

实体	Unweighted			ACE
类型	Pre	Rec	F	Value
FAC	41.0%	37.0%	38.9%	37.0%
GPE	68.5%	71.4%	69.9%	61.3%
LOC	47.3%	50.2%	48.7%	44.9%
ORG	49.8%	58.2%	53.7%	50.1%
PER	47.4%	52.2%	49.7%	49.7%
VEH	38.4%	43.2%	40.7%	13.1%
WEA	43.5%	33.8%	38.1%	25.6%
总体	51.2%	55.3%	53.2%	49.0%

比第五位的 76.7 %高 1.6 %。我们得到的待消解项的识别和分类结果经过 Baseline 进行共指消解后得到的 ACE Value 在 ACE 2007 的 EDR 评测中排在第四位。

3.2　谱聚类结果

与 Baseline 不同的是,谱聚类方法根据最大熵分类模型的输出,即待消解项两两间的共指的概率值,构建邻接矩阵。然后通过谱聚类的方法将指向同一个实体的待消解项聚在一起,从而形成多个实体。在实验中,我们根据待消解项识别和分类的结果,对每一种子类的待消解项构建各自的邻接矩阵。这样做一方面有效地降低了算法的运行时间,另一方面是因为属于不同子类的待消解项一般不会共指一个实体,尤其是对 Diagnostic 的评测。

但是我们需要为每个子类设置一个谱聚类的阈值。这些阈值主要集中在(0.17,0.25)之间。对于平均包含指代项较多的实体类型,如 GPE 的所有子类,PER_Individual 等阈值分别是 0.04 和 0.085;对于平均包含指代项较少的实体类型,如 WEA (Weapon 类型)的所有子类的阈值为 0.45。

和 Baseline 类似,表 6 和表 7 是谱聚类的方法得到的实验结果。通过对比可以发现,谱聚类的方法在召回率(Recall)上略有降低,但有效地提高了准确率(Precision),即能够较好地解决文献[3,5-7]方法中在归并共指链的过程中存在的问题。这符合谱聚类的理论,即谱聚类能够将所有可能存在共指关系的待消解项作为一个整体来看,在每次切分的时候能从全局的角度是切分后的类内距离最小,类间距离最大。并且通过阈值可以控制每个切分后的类中指代项的平均个数。所以,谱聚类能够得到较好的共指消解结果。

表 8 和表 9 是用谱聚类的方法在 Open Test 上得到的结果。我们使用的每个子类的谱聚类阈值与 Diagnostic 中的完全一致。与 Diagnostic 的结果类似,谱聚类方法在 Open Test 上也提高了准确率(Precision)。这说明谱聚类的方法对于实体切分准确率的提高是稳定的。用谱聚类的方法得到的 ACE Value 在 ACE 2007 的 EDR 任务上能排在第三位,但应该注意到,我们的待消解项识别和分类的结果(EMD)比第三位的低 3.8 %,因此,如果提高 Open Test 上待消解项识别和分类的结果,那么谱聚类的共指消解的结果将会更好。

表 6　谱聚类(Diag)按语料结果

语料	Unweighted			ACE
类型	Pre	Rec	F	Value
bn	90.4 %	90.7 %	90.6 %	78.8 %
nw	90.2 %	89.7 %	89.9 %	76.8 %
wl	88.9 %	85.1 %	87.0 %	72.0 %
总体	89.9 %	88.7 %	89.3 %	76.2 %

表 7　谱聚类(Diag)按实体结果

实体	Unweighted			ACE
类型	Pre	Rec	F	Value
FAC	92.6 %	92.8 %	92.7 %	82.4 %
GPE	95.3 %	90.3 %	92.7 %	82.1 %
LOC	93.9 %	93.0 %	93.4 %	86.4 %
ORG	89.1 %	91.1 %	90.1 %	80.0 %
PER	87.2 %	85.7 %	86.4 %	66.7 %
VEH	88.7 %	86.4 %	87.5 %	65.2 %
WEA	88.2 %	90.4 %	89.3 %	79.6 %
总体	89.9 %	88.7 %	89.3 %	76.2 %

表 8　谱聚类(Open Test)按语料结果

语料	Unweighted			ACE
类型	Pre	Rec	F	Value
bn	60.1 %	58.3 %	59.2 %	54.5 %
nw	58.9 %	58.2 %	58.6 %	51.3 %
wl	49.4 %	45.7 %	47.5 %	43.7 %
总体	56.8 %	55.0 %	55.9 %	50.4 %

表 9　谱聚类(Open Test)按实体结果

实体	Unweighted			ACE
类型	Pre	Rec	F	Value
FAC	43.4 %	36.5 %	39.6 %	37.1 %
GPE	76.6 %	71.1 %	73.8 %	62.1 %
LOC	51.8 %	48.6 %	50.2 %	45.7 %
ORG	54.7 %	57.6 %	56.1 %	51.4 %
PER	53.1 %	52.3 %	52.7 %	45.3 %
VEH	49.7 %	42.7 %	45.9 %	31.0 %
WEA	40.9 %	32.8 %	36.4 %	22.2 %
总体	56.8 %	55.0 %	55.9 %	50.4 %

4 总结与展望

本文提出了一种使用谱聚类进行共指消解的方法。该方法能有效地解决常用共指消解方法中在共指链归并时导致的准确率较低的问题。我们通过对每一种子类构建邻接矩阵的办法,降低了算法的运行时间。通过实验表明,基于谱聚类的共指消解算法的评测结果比 Baseline 有较明显的提高,并且该方法对准确率的提高是稳定的。

在接下去的工作中,我们将进一步研究如何能自动地得到每个子类的谱聚类阈值。我们认为,每种子类实体的平均含有的指代项的个数和该子类实体个数的比值可能与谱聚类的阈值存在某种联系。

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] 钱伟,郭以昆,周雅倩,等. 基于最大熵分类模型的英文名词短语指代消解[J]. 计算机研究与发展,2003,40(9):1337-1343.

[3] 庞宁,杨尔弘. 基于最大熵分类模型的共指消解研究[J]. 中文信息学报,2008,22(2):24-27.

[4] 周俊生,黄书剑,陈家骏,等. 一种基于图划分的无监督汉语指代消解算法[J]. 中文信息学报,2007,21(2):77-82.

[5] Yaqian Zhou, Changning Huang, Jianfeng Gao, et al. Transformation Based Chinese Entity Detection and Tracking[C]// IJCNLP, 2005: 232-237.

[6] Wee Meng Soon, Hwee Tou Ng, Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguist, 2001, 27(4): 521-544.

[7] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution[C]// ACL, 2002: 104-111.

[8] Florian, R., Hassan, H., Ittycheriah, A., et al. A statistical model for multilingual entity detection and tracking[C]// NAACL/HLT, 2004: 1-8.

[9] Luo X, Ittycheriah A, Jing H, et al. A mention-synchronous coreference resolution algorithm based on the bell tree[C]// Proc of ACL, 2004: 135-142.

[10] Aron Culotta, Michael Wick, Andrew McCallum, First-Order Probabilistic Models for Coreference Resolution[C]// NAACL/HLT, 2007: 81-88.

[11] AL Berger, VJ Della Pietra, SA Della Pietra. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1): 39-71.

[12] 高琰,谷士文,唐琎,等. 机器学习中谱聚类方法的研究[J]. 计算机科学,2007,34(2):201-203

[13] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Transaction on PAMI, 2000, 22(8): 888-905.

阅读此文的还阅读了：