

基于聚类的中文共指消解全局优化

刘未鹏 周俊生 黄书剑 陈家骏

南京大学计算机科学与技术系

报告人：黄书剑

Outline



- 共指消解问题描述
- 共指消解方法的相关研究
- 基于聚类的中文共指消解全局优化方法
- 实验结果及分析
- 小结

共指消解问题描述

- 指代是自然语言中一种的非常普遍和常见的语言现象,是一个复杂综合的过程。
- 指代一般可分成两种情况: [Wang, 2002]
 - 回指, 是指当前的指示语与上文出现的词、短语或句子(句群) 存在密切的语义关联性;
 - 张三对李老师说, 他感到很不舒服, 不能参加下午的讨论会。
 - 共指, 则主要是指多个名词(包括代名词、名词短语) 指向真实世界中的同一参照体。
 - 共指关系是等价关系, 可以独立于上下文存在。而回指不一定满足等价性原则。
 - “香港首任行政长官” 和 “董建华”

共指消解方法的相关研究

- 当前，大多数基于机器学习的共指消解方法一般分为两个步骤[Soon, et al., 2001]。
 - 第一步：分类
 - 使用统计机器学习方法对Mention Pairs 进行二元分类；
 - 第二步：聚类(*本文关注的内容)
 - 在得到了Mention Pair的共指概率的基础上构造最佳的共指链。
 - BestCut方法 [C. Nicolae and G. Nicolae, 2006]
 - 整数线性规划方法 [Finkel and Manning, 2008]
 - Bell树模型 [Luo et al., 2004]

共指消解方法的相关研究（续）

- 存在的问题：
 - 损失函数的指定比较随意。
 - 最大化类内相似度，最小化类间相似度
 - 最终将全部聚集到同一个类中
 - BestCut: 寻找经过边的权值和最小的cut
 - 需要SVM来进行cut停止条件判断
 - 使用的最优化模型和算法并没有充分结合共指消解问题本身的特质。

一种基于最小化决策错误的损失函数

- 假设1: 二元分类的准确程度相对较高, 因而全局聚类优化方法则应该更加关注于对违反分类决策的聚类进行惩罚。

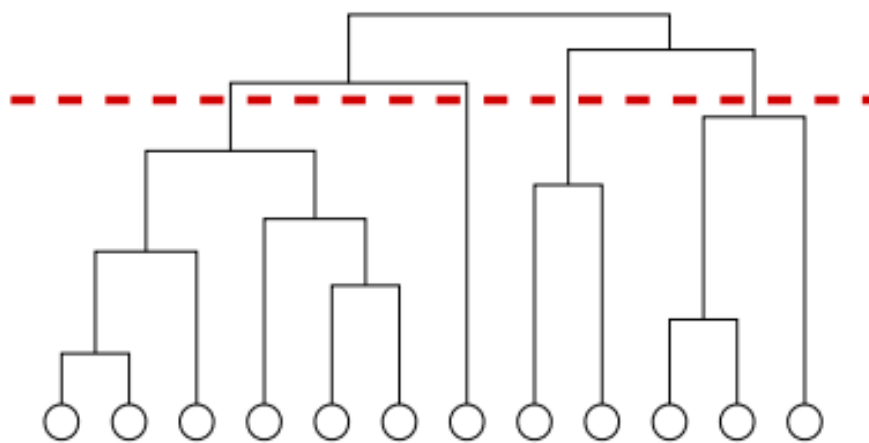
The diagram shows the loss function with two callouts. The first callout, labeled '类间共指的概率之和' (Sum of co-mention probabilities between clusters), points to the first summation term. The second callout, labeled '类内非共指的概率之和' (Sum of non-co-mention probabilities within clusters), points to the second summation term.

$$\operatorname{argmin} \sum_{i,j;i \neq j} \sum_{e_{ij} \in C_{ij}; P(e_{ij}) > 0.5} P(e_{ij}) + \sum_i \sum_{e_{ii} \in C_{ii}; P(e_{ii}) < 0.5} \bar{P}(e_{ii})$$

- 其中, C_{ij} 表示聚类*i*和聚类*j*之间所有的边的集合, C_{ii} 表示聚类*i*内部的所有的边的集合, e_{ij} 和 e_{ii} 则表示上述集合中的某一条边; $P(e)$ 表示某条*e*的权重, 即两个Mention之间的共指概率, $\bar{P}(e) = 1 - P(e)$

自底向上的聚类

- 算法描述：
 - ▣ 初始化：每个 Mention 作为一个单个的聚类。
 - ▣ 自底向上地不断寻找能够使得目标函数递降最多的合并。
 - ▣ 在得到的聚类层次中找到那个使得目标函数最小的聚类层次。



利用局部信息的聚类

- 传统基于二元分类的共指消解具有信息匮乏性，Mention Pair携带的上下文信息往往是不够的；如前文所述的全局优化方法往往代价较大，效率不够高。
- 为此我们提出了两种利用部分全局信息的聚类方法，这两种方法都基于这样一个假设：

假设2： 一个Mention Pair携带的信息可能不足，而能够对其进行补充的信息则存储于跟它共指概率较高的周边Mention中。

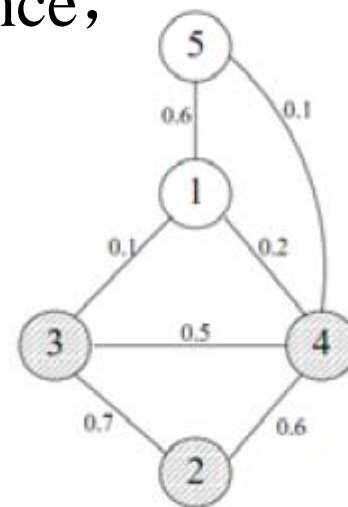
凝聚式的Family-Vote

- 选择共指概率最高的一对Mention作为初始聚类。
- 在其余所有Mention中寻找 Acceptance 最高的Mention加入这个聚类。
- 重复这个过程。直到再也找不到 $\text{Acceptance} > 0$ 的Mention，该Cluster便构造结束。
- 重复刚才的步骤构造剩下的Clusters。
- 算法中的 $\text{Acceptance} = \text{CostOfNotLinking} - \text{CostOfLinking}$

- 其中，
$$\text{CostOfNotLinking} = - \sum_{m_i \in C} \log \bar{p}(i, k)$$
$$\text{CostOfLinking} = - \sum_{m_i \in C} \log p(i, k)$$

凝聚式的Family-Vote (例)

- $Mary_1$ has a brother₂ John₃. The boy₄ is older than the girl₅.
 - ▣ Step1: 选取置信度最高的2-3 pair作为起始;
 - ▣ Step2: 依次判断其余各点的acceptance, 并依此将4加入共指链中;
 - ▣ Step3: 完成该共指链的构造
 - ▣ Step4: 选取剩余节点中置信度最高的1-5 pair, 重复上一过程。



分裂式的Family-Vote

- 对每一个 $\langle \text{Mention}_i, \text{Mention}_j \rangle$ 对，记与 Mention_i 一阶共指的所有 Mention 集合为 S_i ，与 Mention_j 一阶共指的所有 Mention 集合为 S_j ，令：

$$\text{CostOfNotLinking} = - \sum \log \bar{P}(p, q)$$

$$\text{CostOfLinking} = - \sum_{m_p \in S_i; m_q \in S_j} \log P(p, q)$$

- 如前述计算 Acceptance，若 $\text{Acceptance} < 0$ ，则切断 Mention_i 和 Mention_j 之间的边，否则保留这条边。
- 对经过上一步处理之后的图求出各连通分量。

分裂式的Family-Vote (续)

- 一阶共指指的是非传递关系的共指
 - ▣ 一阶共指集合越大，获得的周围信息就越多，相应的噪声也会变大；
 - ▣ 反之，获得周围的信息就越少，但相对较为精确。
- 保守的识别策略：
 - ▣ 仅认为共指概率大于0.8的Mention之间存在一阶共指关系。
 - ▣ 考虑到如果一阶共指集合中含有被错误划分为共指的Mention，会使得CostOfLinking增大较多，从而影响共指的判断，

实验设计

- 实验数据
 - ACE 2005评测的训练数据
- 分类算法
 - C4.5算法
- 评分方法
 - MUC6 score
- Baseline系统
 - 最近优先策略(link-first)
 - 最佳优先策略(link-best)

特征选择



- 距离特征
- i-代词特征
- j-代词特征
- 字符串匹配特征
- 指示性的Mention特征
- 单复数一致性特征
- 语义类别一致性特征
- 性别一致性特征
- 专有名词特征
- 同位语特征

实验结果与分析

Approaches	Precision(%)	Recall(%)	F(%)
link-first	77.56	75.54	76.54
link-best	78.42	75.36	76.86
Bottom-up	77.91	78.06	77.98
Family-Vote(1)	77.59	78.01	77.80
Family-Vote(2)	77.60	77.89	77.74

- 三种方法与Baseline 相比都有1%左右的提高；
- 基于最小化决策错误的损失函数在聚类时综合考虑了全局的信息，取得了最好的效果；
- 两种Family-Vote的方法虽然主要利用的只是部分全局信息，但和最好的结果相差并不明显。

小结

- 本文主要针对指代消解中的聚类过程进行了分析和讨论，力图在聚类过程中有效的利用全局或局部的信息，从而使得聚类更加准确合理。
- 本文提出了一种最小化决策错误的聚类损失函数，以及两种基于部分全局信息的聚类策略。实验表明，利用全局信息确实能对聚类结果产生有益的影响。
- 今后的工作中我们更全面对对全局信息的使用进行研究和探讨

主要参考文献

- [Soon, et al., 2001] W. M. Soon, H. T. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.
- [C. Nicolae and G. Nicolae. 2006] C. Nicolae and G. Nicolae. 2006. Bestcut: A graph algorithm for coreference resolution. In: D. Jurafsky and E. Gaussier eds. *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia: Association for Computational Linguistics, 275-283.
- [Finkel and Manning, 2008] Jenny Rose Finkel and Christopher D. Manning. 2008. Enforcing Transitivity in Coreference Resolution. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008), Short Papers*, pp. 45-48.
- [Luo, et al., 2004] X. Luo, et al. 2004. A mention-synchronous coreference resolution algorithm based on the bell tree. In: D. Scott ed. *Proc. of the 42th annual meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, 135-142.
- [Newman and Girvan, 2004] Newman, MEJ. Fast algorithm for detecting community structure in networks. *Phys Rev E*. 2004;69(no 066133)
- [Vilain, et al., 1995] M Vilain , J Aberdeen et al. A model theoretic coreference scoring scheme. In : *Proc. of the 6th Message Understanding Conf (MUC6)* , San Francisco : Morgan Kaufmann Publishers, 1995, 45-52.
- [Yang, et al., 2004] Yang, X., G. Zhou, J. Su, and C. L. Tan. An NP-Cluster Based Approach to Coreference Resolution. *Proceedings of the 20th International Conference on Computational Linguistics*, 2004, Geneva, Switzerland, pp. 226-232.