第13章：篇章分析

课件制作：王新宇、吴昊一、屠可伟
讲解人：王新宇

篇章的定义

- **篇章**是由一个句子序列构成的连贯结构。
- 是什么使段落连贯？
  - 连贯性体现在句子或短语之间有意义的关系。

例子

- Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term. The impact won't be that great, said Graeme Lidgerwood of First Boston Corp. This is in part because of the effect of having to average the number of shares outstanding, she said. In addition, Mrs. Lidgerwood said, Norfolk is likely to draw down its cash initially to finance the purchases and thus forfeit some interest income.

词汇链

- Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term. The impact won't be that great, said Graeme Lidgerwood of First Boston Corp. This is in part because of the effect of having to average the number of shares outstanding, she said. In addition, Mrs. Lidgerwood said, Norfolk is likely to draw down its cash initially to finance the purchases and thus forfeit some interest income.

共指链

- Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term. The impact won't be that great, said Graeme Lidgerwood of First Boston Corp. This is in part because of the effect of having to average the number of shares outstanding, she said. In addition, Mrs. Lidgerwood said, Norfolk is likely to draw down its cash initially to finance the purchases and thus forfeit some interest income.

篇章标记

- Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term. The impact won't be that great, said Graeme Lidgerwood of First Boston Corp. This is in part because of the effect of having to average the number of shares outstanding, she said. In addition, Mrs. Lidgerwood said, Norfolk is likely to draw down its cash initially to finance the purchases and thus forfeit some interest income.

篇章衔接的特征

- 篇章衔接的特征主要分为三类
  - 词汇重叠/词汇链
  - 共指链 ← 有待详细讨论
  - 提示词/篇章标记

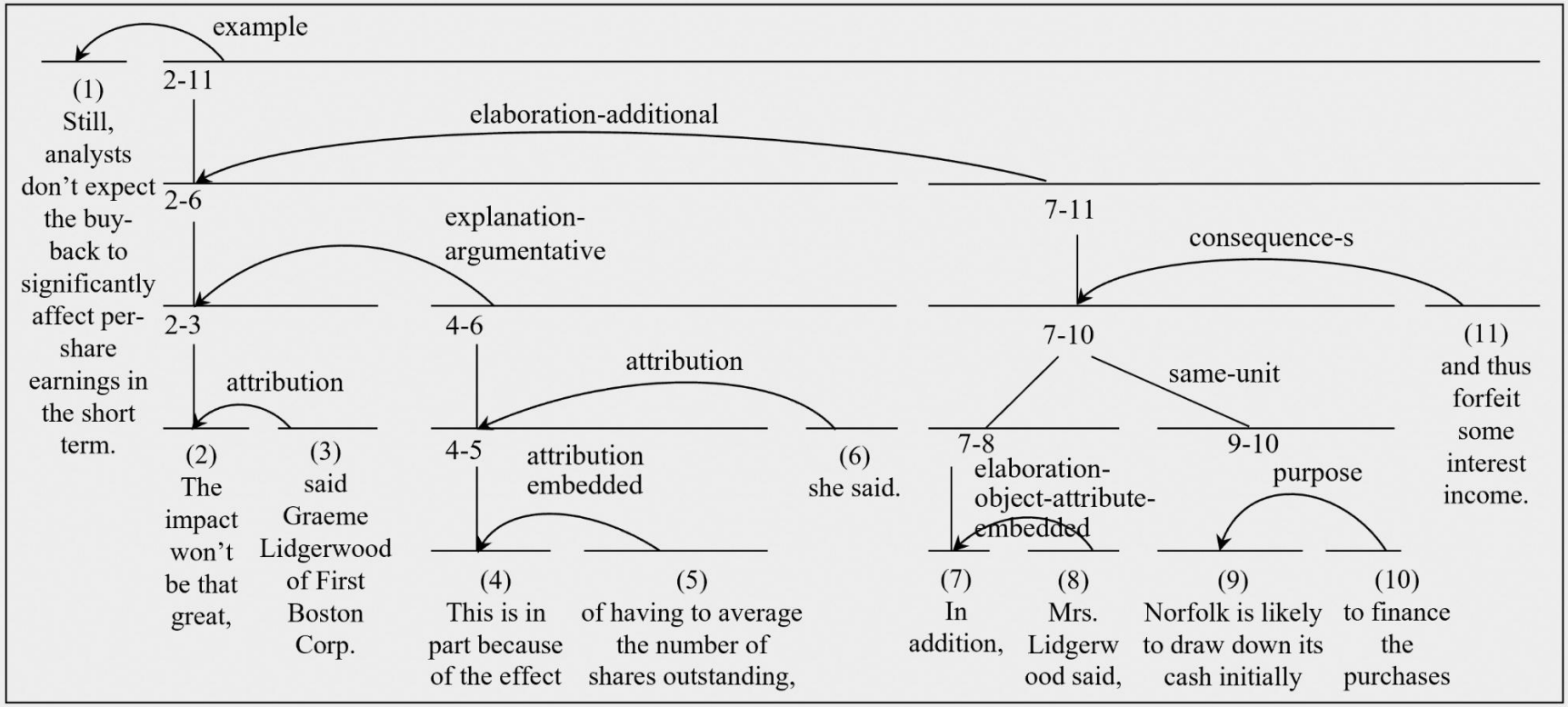
连贯性关系

- 文本跨度之间的连接可以指定为一组连贯性关系。
- 修辞结构理论（RST）
  - 最常用的一致性关系模型
  - 大多数关系都存在于核心和卫星之间。
    - 核心：作者目的的核心，可独立解释
    - 卫星：不太中心，通常只能根据核心来解释

关系示例

- **原因**
  - [$_{NUC}$ 小红从上海坐火车到北京。] [$_{SAT}$她要去参加一个会议。]
- **阐述**
  - [$_{NUC}$小红来自内蒙古。] [$_{SAT}$她住在内蒙古大草原的中间。]
- **证据**
  - [$_{NUC}$ 小明一定在这里。] [$_{SAT}$他的车停在外面。]
- **归因**
  - [$_{SAT}$分析师估计表示] [$_{NUC}$本季度美国商店的销售额也有所下降]
- **列表**
  - [$_{NUC}$ 小芬是他的伴侣；] [$_{NUC}$浩然从事商业]

层级篇章结构



不对称关系用从卫星到核心的弯曲箭头表示。

篇章分析

- 给定一系列句子，确定它们之间的连贯关系
- 两个阶段：
  - 基本篇章单元分割分割：将句子分割成一系列基本篇章单元
    - 可以作为序列标注进行分析
  - 篇章结构分析：在基本篇章单元上构建篇章结构
    - 可以转换为成分或依存分析

共指消解

ElitesAI

什么是共指消解

- 识别所有提及同一实体的提及
- 例子：
  - Still, analysts don't expect the buy-back to significantly affect per-share earnings in the short term. The impact won't be that great, said Graeme Lidgerwood of First Boston Corp. This is in part because of the effect of having to average the number of shares outstanding, she said. In addition, Mrs. Lidgerwood said, Norfolk is likely to draw down its cash initially to finance the purchases and thus forfeit some interest income.

两步共指消解

- 1. 检测提及（简单）
  - “[我]投票给[小明]，因为[他]最符合[[我的]价值观]，”[她]说
  - 提及可以嵌套！
- 2. 对提及进行聚类（困难）
  - “[我]投票给[小明]，因为[他]最符合[[我的]价值观]，”[她]说

提及类型

- 名词短语
  - *一篇文章*、*那本书*、*小明*定名词短语
- 人称代词
  - 我、它、it、they
- 指示代词
  - 这、那个、this、these
- 零形回指
  - 我昨天生病了，没去上课。 【0】今天好转了。

提及检测

- 使用词性标注器 + 跨度分析器 + 命名实体标注器来检测不同类型的名词短语和代词
- 但还需要进一步过滤
  - 许多名词短语不是提及
    - 小明没电脑。
    - 南河二是全天最亮的恒星之一。
  - 甚至代词也可能不是提及
    - It ‘s sunny.
    - As we know…
    - It is essential…
- 方法
  - 规则
  - 二元分类
  - 与提及聚类的联合推理

提及聚类

*“**我**投票给**小明**，因为**他**最符合**我的**价值观，”**她**说。*

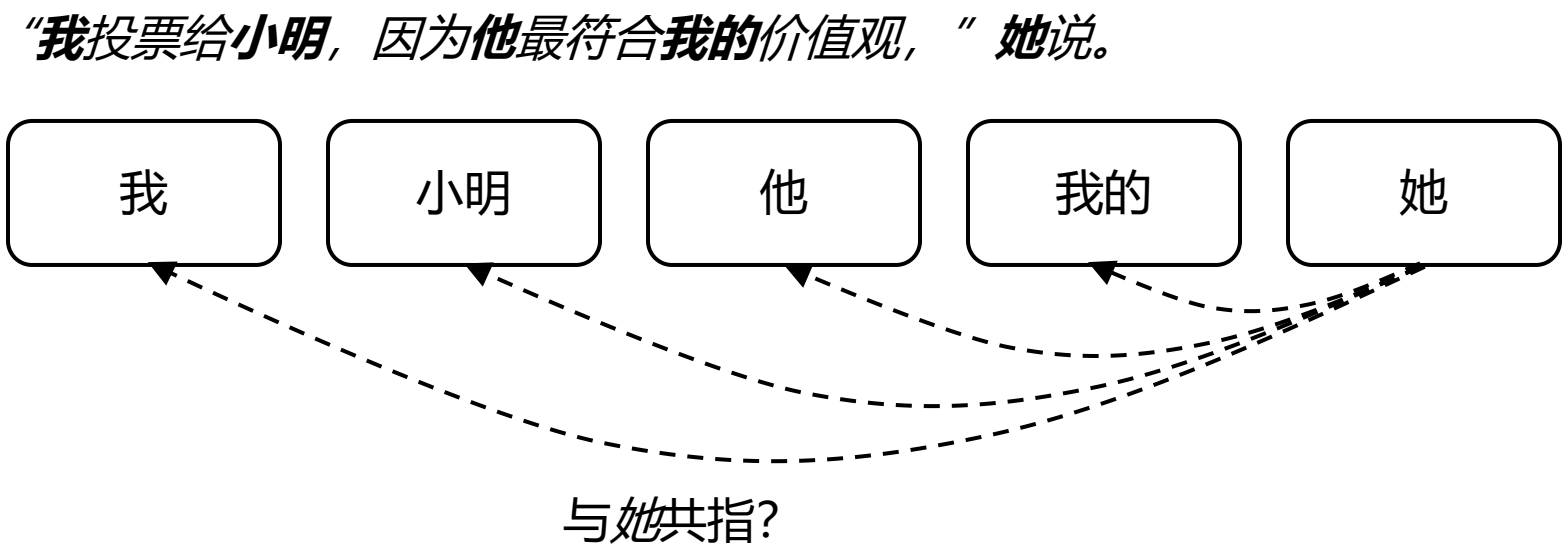
我	小明	他	我的	她

提及聚类

*“**我**投票给**小明**，因为**他**最符合**我的**价值观，”**她**说。*

提及聚类

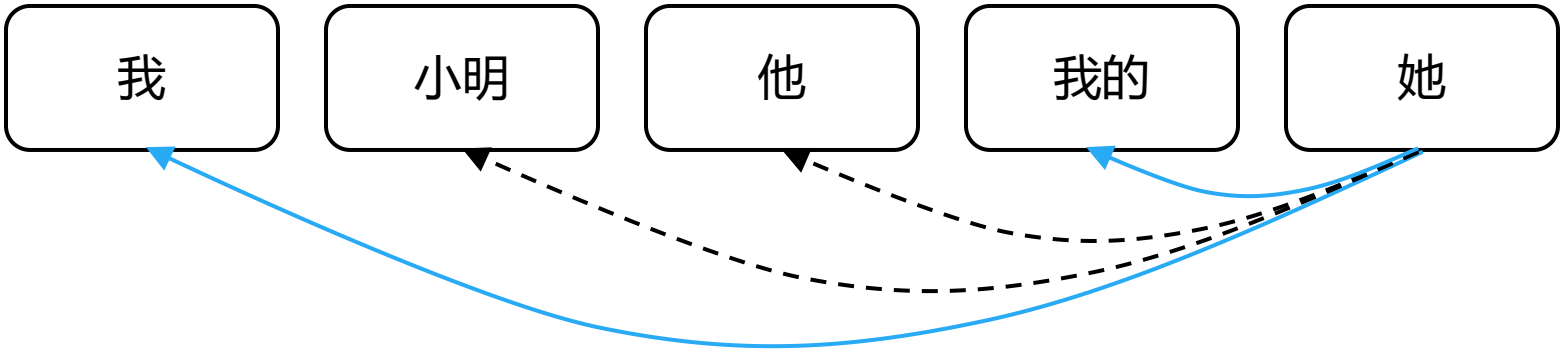
- 训练一个二元分类器，为每对提及分配一个共指概率
  - 例如：对于“她”，查看所有候选先行词（先前出现的提及）并确定哪些是其共指词



提及聚类

- 训练一个二元分类器，为每对提及分配一个共指概率
  - 例如：对于“她”，查看所有候选先行词（先前出现的提及）并确定哪些是其共指词

*“**我**投票给**小明**，因为**他**最符合**我的**价值观，”**她**说。*

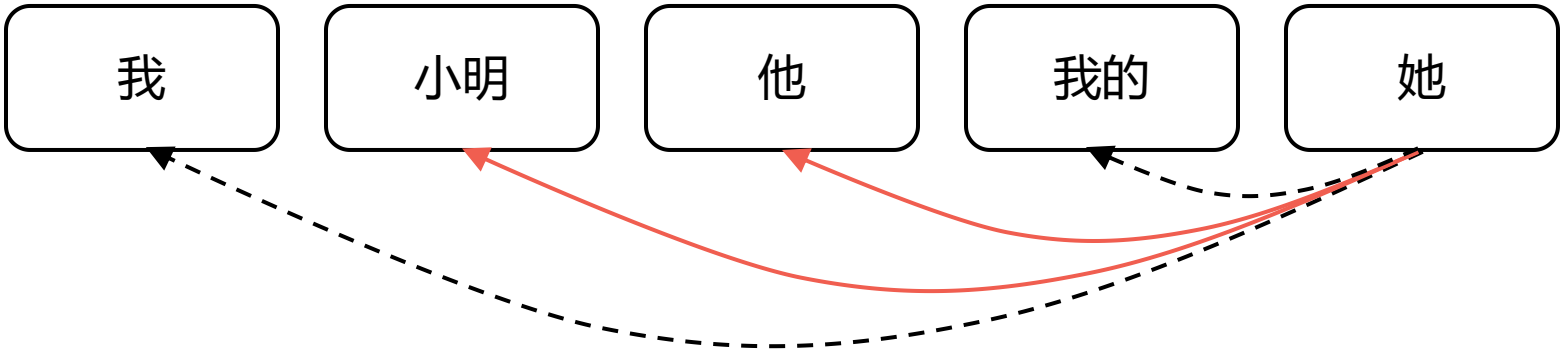


正样本：希望$q(m_i, m_j)$ 接近 1

提及聚类

- 训练一个二元分类器，为每对提及分配一个共指概率
  - 例如：对于“她”，查看所有候选先行词（先前出现的提及）并确定哪些是其共指词

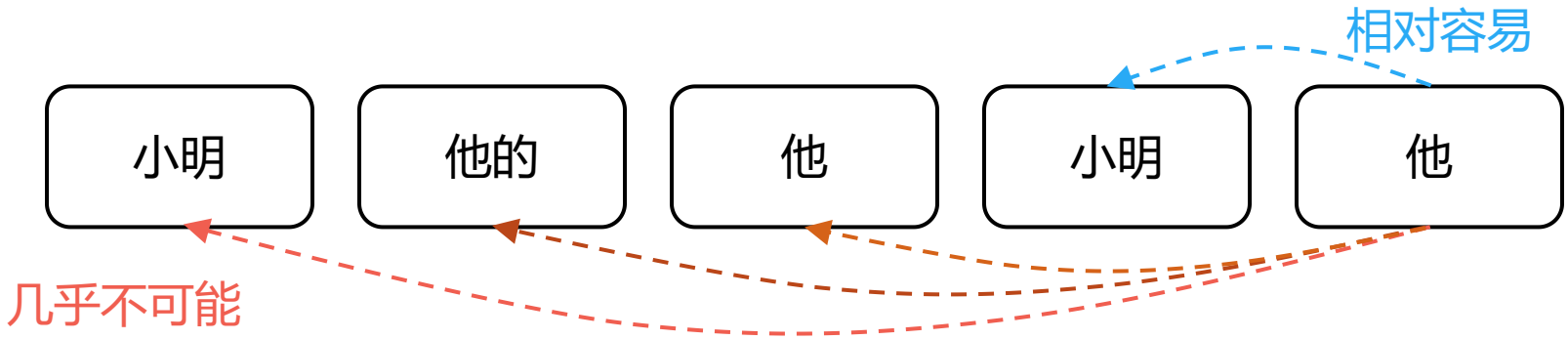
*“**我**投票给**小明**，因为**他**最符合**我的**价值观，”**她**说。*



负样本：想要$q(m_i, m_j)$ 接近 0

提及聚类：缺点

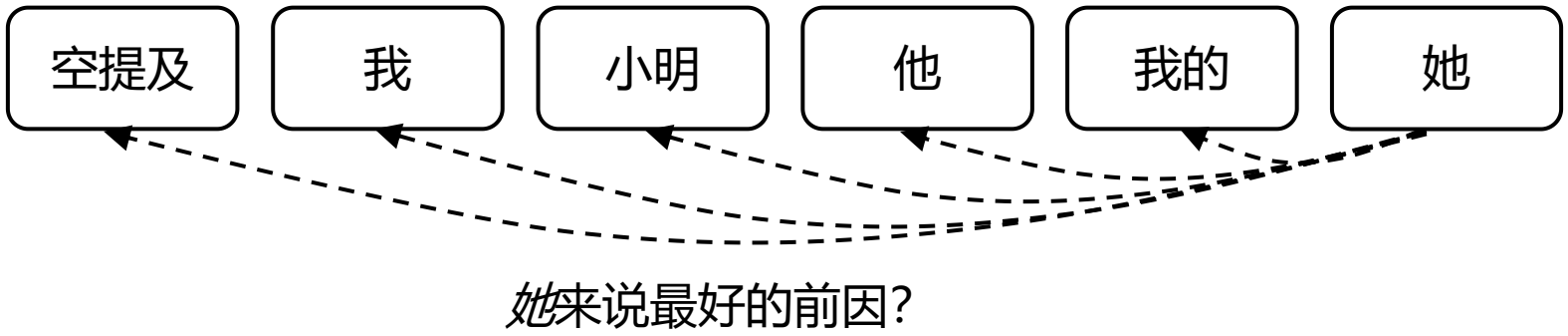
- 假设我们有一个很长的文档，其中提到了以下内容
  - **小明……他的……他……**〈几段〉……*投票给**小明**，因为**他**……*



- 许多提及只有一个明确的前因
  - 但我们要求模型预测所有这些
- 解决方案：训练模型以仅预测每次提及的一个先行词

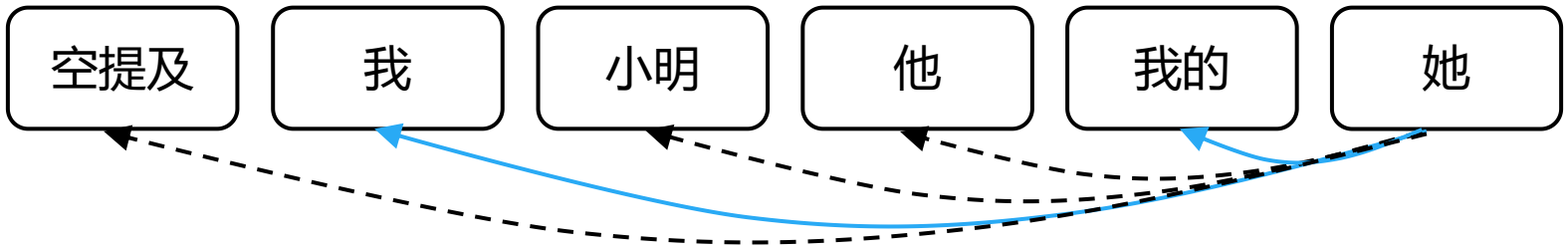
共指模型：提及排名

- 根据模型为每个提及分配其得分最高的候选先行词
  - 空提及允许模型拒绝将当前提及与任何内容相关联（“单一”或“首次”提及，或不提及）



共指模型：提及排名

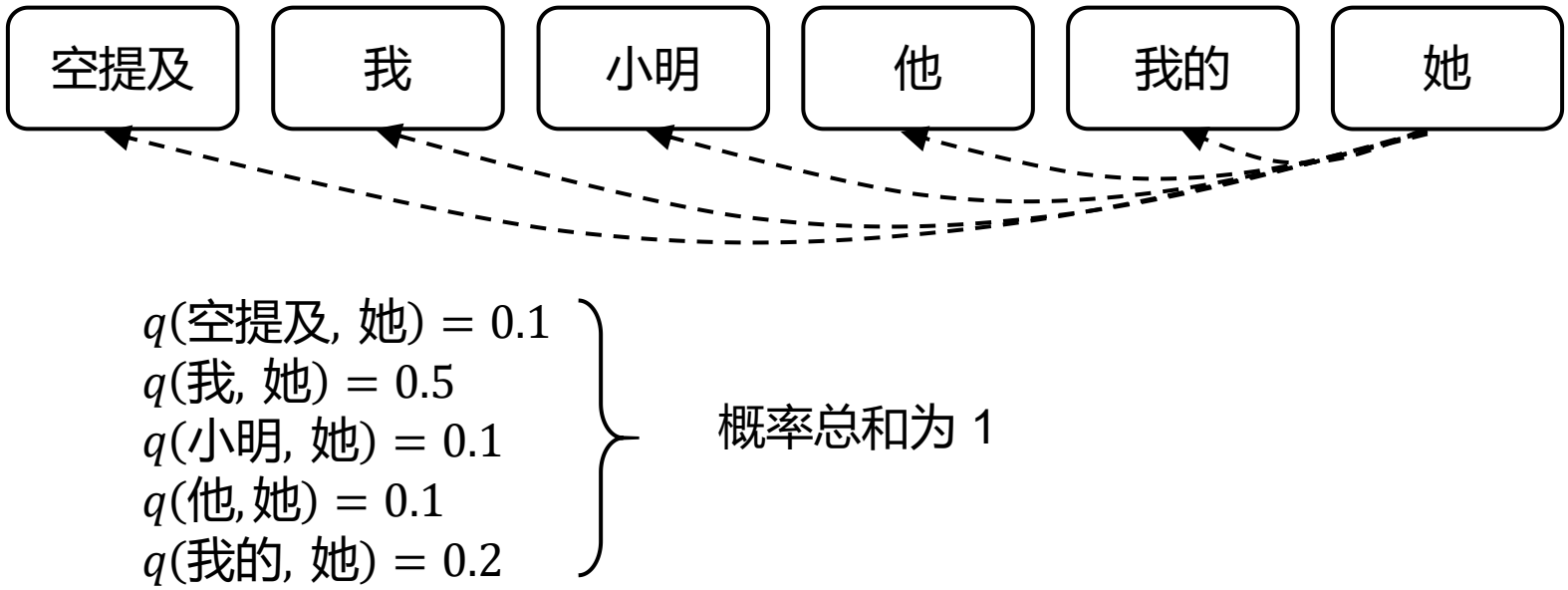
- 根据模型为每个提及分配其得分最高的候选先行词
  - 空提及允许模型拒绝将当前提及与任何内容相关联（“单一”或“首次”提及，或不提及）



正样本：模型必须为其中之一分配高概率（但不一定是两者）

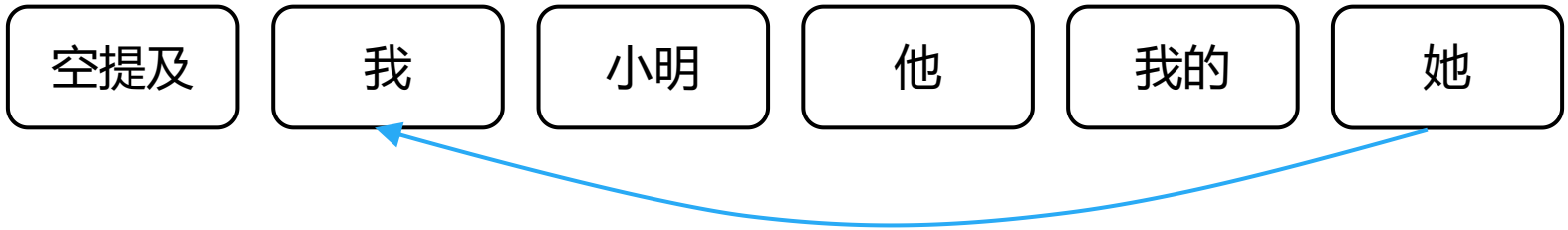
共指模型：提及排名

- 根据模型为每个提及分配其得分最高的候选先行词
  - 空提及允许模型拒绝将当前提及与任何内容相关联（“单一”或“首次”提及，或不提及）



共指模型：提及排名

- 根据模型为每个提及分配其得分最高的候选先行词
  - 空提及允许模型拒绝将当前提及与任何内容相关联（“单一”或“首次”提及，或不提及）



q(空提及, 她) = 0.1
q(我, 她) = 0.5
q(小明, 她) = 0.1
q(他, 她) = 0.1
q(我的, 她) = 0.2

概率总和为 1

共指模型：训练

- 我们想要当前的提及m_i 链接到与其共指的*任*一候选先行词。
- 从数学上来说，我们希望最大化这个概率：

$$\sum_{j=1}^{i-1} \mathbf{1}(y_{ij} = 1) \cdot \frac{\exp(s(m_i, m_j))}{\sum_{j'=1}^{i-1} \exp(s(m_i, m_{j'}))}$$

迭代所有候选词之前的（之前出现的提及）

对于那些是共指的到m_i…

……我们希望模型能够分配一个高概率

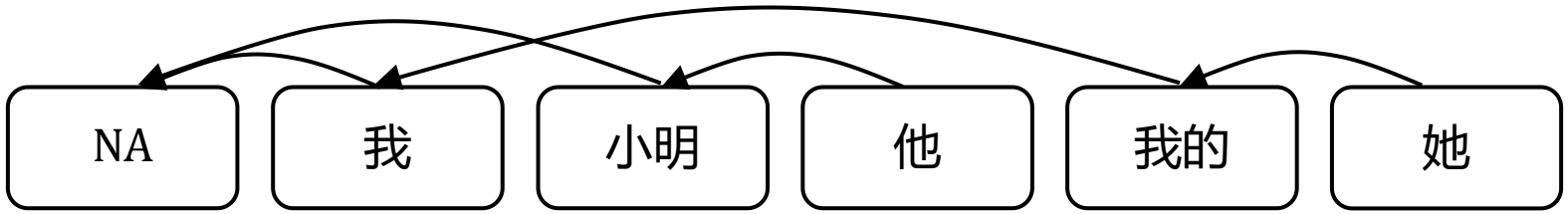
计算概率

- …根据特征
  - 人员/人数/性别协议
    - 杰克送给玛丽一份礼物。她很兴奋。
  - 语义兼容性
    - ……矿业集团……公司……
  - 某些语法限制
    - 约翰给他买了一辆新车。［他不可能是约翰］
  - 最近提到的实体首选参考
    - 约翰去看电影了。杰克也去了。他不忙。
  - 语法角色：更喜欢主语位置上的实体
    - 小明和小红一起去看电影。他不忙。
  - 并行性
    - 小红和小明去看电影。小张和他一起去了一家酒吧。
- 或者只使用神经网络

推理

- 对每个前因的前因分数进行排名。最高者获胜。

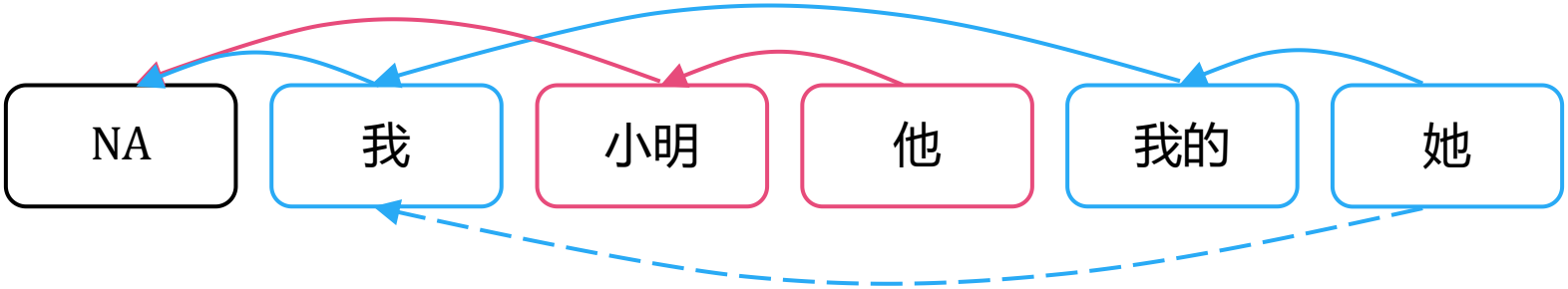
*“**我**投票给**小明**，因为**他**最符合**我的**价值观，”**她**说。*



推理

- 对每个前因的前因分数进行排名。最高者获胜。
- 采取传递闭包来得到聚类
  - 不包括北美

*"**我**投票给**小明**，因为**他**最符合**我的**价值观，"**她**说。*



尽管模型没有预测这种共指链接，但由于传递性，

总结

ElitesAI

篇章分析

- 篇章是一组连贯的结构化句子。
  - 文本跨度通过连贯关系连接。
  - 这些关系形成了层次结构。
  - 篇章分析：EDU切分+RST分析

- 共指消解
  - 提及检测
  - 提及聚类
    - 二元分类与排名