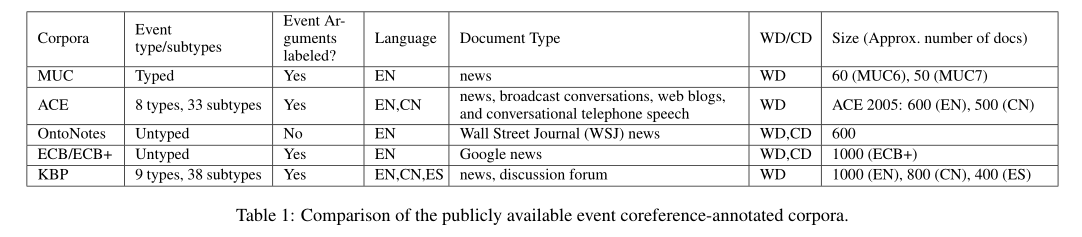
(2018 IJCAI)Event Coreference Resolution: A Survey of Two Decades of Research

时间：2021/8/16

1. **语料库**

从6个维度评估5个语料库

1. 事件(子)类型是否被标注；
2. 事件参数是否有注释；
3. 文件是否为英文(EN)、中文(CN)和/或西班牙文(ES)；
4. 文件类型；
5. 文档是否标注了文档内(WD)和/或跨文档(CD)事件共指链；
6. 文档数量的大致大小。



* MUC

MUC没有正式定义/评估事件共指：事件共指是(隐含地)需要作为场景模板填充任务的一部分来执行的任务。

* ACE

在ACE中，事件被定义为“一种特定事件的发生，通常是一种状态的改变，涉及到参与者”

* OntoNotes

它提供了文档内和跨文档的实体和事件共指注释。然而，它没有指定哪些提及是实体提及，哪些是事件提及，也没有注释所有事件共指链:当且仅当一个链的至少一个事件提及是名义上的时，该链才被注释。

* ECB

事件被描述为“发生或正在发生的情况”（“situations that happen or occur”）。事件可以表示为准时的、持续的或静态的谓词，描述“某物获得或保持为真的状态或环境”。ECB包含属于43种事件类型之一的内部和跨文档事件引用链接，文档内链接只被部分注释。ECB+语料库加入了更多带注释的文档，并用新的注释样式重新注释了现有文档。它还通过将事件建模为四个参数的组合来扩展定义，即动作、时间、位置和参与者。

* KBP

KBP语料库定义了一个复杂的分层事件结构，它超越了任何现有的(和部分重叠的)语料库。它们遵循ACE语料库中的定义，但在几个领域扩展了可标记性：轻微扩展的事件本体，添加通用和其他(非realis)事件提及，为事件提及添加无参数触发词，为联系和事务事件添加附加属性，为多种类型/子类型添加事件提及的双重标记，以及为某些协调类型添加事件提及的多重标记。

1. **模型**

**监督模型**

* 提及对模型（Mention-Pair Models）

许多事件共指解析器采用两步解析框架。

在第一步中，使用二进制分类器(称为提及对模型)来确定两个事件提及是否相关。提对模型通常使用现成的学习算法进行训练（决策树，最大熵，支持向量机和深度神经网络）。

在第二步中，需要一个单独的聚类机制来协调成对决策并构建一个分区。一部分人采用**凝聚聚类算法**（如最近优先聚类和最佳优先聚类）；其他人使用**图划分**（给定一个测试文档，首先构建一个无向加权图，其中节点代表文档中提到的事件，边的权重代表它所连接的两个节点的共指可能性，然后利用谱聚类和分裂聚类等聚类算法获得共指聚类）

**对该方法的改进**：包括使用特征加权来训练更好的模型和训练多个分类器来处理不同句法类型的事件提及之间的共指。

* 生成模型（Generative Models）

提及对模型和相关的两步方法受到误差传播的影响，其中提及对模型产生的误差会传播到聚类步骤。为了解决提及对模型的误差传播，yang等人[2015]提出了一个**监督的非参数生成模型**用于事件共指解析。作为一种聚类模型，事件提及直接分配给增量构建的共指聚类。作为一个非参数模型，聚类的数量不需要预先知道。作为贝叶斯模型，它可以利用先验信息，在这种情况下，先验信息编码由提及对模型提供的知识。最后，该模型被监督，在建模过程中可以使用丰富的特征。

* 提及-排名模型（Mention-Ranking Models）

提及对模型认为事件提及的每个候选先行词由独立于其他候选先行词进行解析，只能确定一个候选先行词相对于所提到的事件有多好，但不能确定它相对于其他候选先行词有多好。排名模型（Ranking models）通过允许同时对提及的候选先行词进行排名来解决这一问题。

受其成功应用于实体共指消解的激励，Lu和Ng 训练了一个**概率提及-排名模型**[2017]，该模型对事件提及的候选先行词进行排名，使其正确的先行词具有最高的排名。

* Easy-First模型（Easy-First Models）

简单优先的共指模型以一种迭代的方式运行，其目标是先做出简单的链接决策，然后利用这些简单决策(作为额外的知识)来做出艰难的链接决策。

最早采用简单优先方法的事件共指解析器之一是**Stanford的解析器**[2012]。此解析器使用实体共指输出迭代地引导事件共指输出，反之亦然。

Liu等人[2014]试图通过添加**第三步**改进上述两步“分类和聚类”方法，在第三步中，他们不断地将事件共指聚类中的一个提到的参数传播到同一聚类中的另一个提到的参数，直到事件共引用聚类中的所有提到都共享相同的参数。

Lu和Ng[2016]使用6个**筛选器**实现了一种简单优先的事件共指解析方法。筛选器由一组手工规则或机器学习的分类器组成，用于对测试文档中提及对的子集进行分类。六个筛选器按精确度递减的顺序排列成一个管道。当两个事件的提及被筛选器假定共指时，其中一个提及所提取的任何参数将被另一个提及所共享。

Choubey和Huang[2017]构建了一种**两步凝聚聚类算法**，用于文档内和跨文档的事件共指解析。在第一步中，使用在文档内和跨文档事件共指之间交互的迭代算法，来合并文档内或跨文档集群。当不能再进行合并时，该算法进入第二步，在该步骤中，以如下迭代方式合并其他集群：如果集群c1中的提及与c3中的提及紧密相关（即，具有相同的依赖关系）或松散关联（即，在同一个句子语境中），并且c2中的提及也与c3中的提及紧密或松散地联系在一起，那么c1和c2将合并。

* 联合模型（Joint Models）

上述模型都采用管道体系结构，其中在事件共指解析之前提取事件触发词和参数。因此，来自上游组件的错误将传播到事件共指解析器。

错误传播问题的一个解决方案是对IE管道中不同任务的输出进行**联合推理**。联合推理允许事件共指解析器及其上游组件通过利用以人工指定的任务约束形式表示的背景知识来相互影响(并可能改进)。Chen和Ng[2016]通过整数线性规划（ILP）对IE管道中四个关键任务（即实体提取、实体共指、事件提取和事件共指）的模型输出进行联合推理。Lu等人[2016]使用马尔可夫逻辑网络（MLN）在四个任务上（触发器识别、参数提取、实体共指和事件共指）执行联合推理。ILP通常用于编码硬约束，MLNs允许对软规则和硬规则进行编码。

另一个解决方案是**联合学习**。Araki和Mitamura[2015]将联合学习事件触发词检测和事件共指解析的任务形式化为使用结构化感知器训练算法学习的结构化预测问题，采用基于段（segment-based）的解码和多段搜索（multiple-beam search）来识别事件触发器，并将其与最佳优先聚类相结合，用于文档级联合解码中的事件共引用解析。Lu和Ng[2017]共同学习事件共指消解、事件触发检测和事件回指性确定。Lu和Ng构建了一个结构化条件随机场模型，该模型包含（1）一元因子，用于编码每个任务的特定特征；（2）高阶因子，用于以软方式捕获每对任务之间的交互。每个事件提及都与三个输出变量相关联，用于编码其触发词子类型、回指性和先行词。目标是了解这些输出变量的哪个值组合是最可能的。

**半监督模型**

监督模型受到数据采集瓶颈的影响，手动注释IE管道中所有组件的数据非常昂贵。为了解决这个问题，研究人员使用**主动学习**来选择有信息的实例，表明只有少量的训练句子需要注释来实现最先进的事件共指消解性能。另一个尝试是**利用大量的域外文本数据**，其思想是：(1)用五个事件语义成分（动作、论点、时间、地点和句子/子句）来表示事件结构；(2)采用显式语义分析、Brown聚类、Word2Vec和基于依存关系的词嵌入等方法，将每个事件分量转换为对应的向量表示；(3)连接所有组件以形成结构化向量表示。这样，事件共指解析转换为比较事件向量相似性的任务。

**无监督模型**

无监督模型是为消除模型对标注数据的依赖而提出的。现有的无监督事件共指模型绝大多数是概率生成模型。

Bejan和Harabagiu [2014] (B&H)提出了几个**非参数贝叶斯模型**用于事件共指解析，这些模型可以概率推断文档内和跨多个文档的事件集群。模型的优点是可以自动推断文档中事件集群的数量，但模型在表示特征丰富的对象方面仍有局限性。B&H扩展了这个模型，考虑从WordNet和FrameNet派生的额外语言特征，使用丰富的特征表示可能会增加贝叶斯模型的复杂性，并且不能保证所有的特征都对任务有积极的影响。B&H用一个特征选择机制扩展了他们的模型，自动选择一个有限的显著特征。此外，他们还提出了另一种贝叶斯模型，该模型具有捕获对象之间结构依赖关系的机制。

B&H模型利用了从WordNet和FrameNet中提取的语义信息，显著提高了共引用性能。但是，由于缺乏可比较的词汇知识库使得英语以外的语言中事件共指解析器的设计复杂化。为此，Chen和Ng[2015]设计了一个**概率模型**，其参数是使用EM估计的，用于计算两个事件提及共指的概率。它的生成过程不依赖于语言，也不依赖于从词汇知识库中提取的特征，因此它可以应用于既没有注释数据也没有大规模知识库的语言。

1. **语言特征**

**词汇特征（Lexical features）**

词汇特征可以显式或隐式地比较一对事件提及的事件触发词。常用的词汇特征包括

1. 成对特征，如两个事件触发词对和词性对等；
2. 字符串匹配特征，如精确字符串匹配、部分字符串匹配、词干匹配和引理匹配；
3. 触发相似特征，如Dice系数、编辑距离、Jaro系数，以及基于项频向量计算的相似度；
4. 事件触发词的周边词及其相似性。

**参数特征（Argument features）**

参数特征也被广泛用于事件共指消解，因为具有不兼容参数的事件提及不太可能是共指的。

**一类参数特征**编码了两个事件提及之间重叠参数的数量，每个事件提及拥有的唯一参数的数量，以及两个事件提及是否有冲突的时间和地点参数。**其他参数特征**编码参数之间的相似性，如两个参数是否（实体）共指、使用Dice系数的表面相似性以及参数头词之间的Wuparmer WordNet相似性。

要提取参数特性，通常需要一个事件参数提取器和一个实体共引用解析器。虽然在早期的事件共引用工作中使用了黄金参数，但最近的工作侧重于使用自动提取的参数构建端到端解析器。

**语义特征（Semantic features）**

语义特征从词汇语义资源（如WordNet、FrameNet、VerbOcean）、Brown聚类和Word2Vec生成的词嵌入中提取，用于计算两个事件提及之间的相似度。基于嵌入的相似度特征在所有特征中权重最高，事件(子)类型匹配也被证明是事件共指的有力指标。

**语篇特征（discourse features）**

语篇特征重新编码两个事件提及之间的标记、事件和句子距离，以及一个事件提及在文章中的位置。消融实验表明，语篇特征对事件共参性能的贡献不如其他类型的特征。

1. **评估**

**提取候选事件提及**

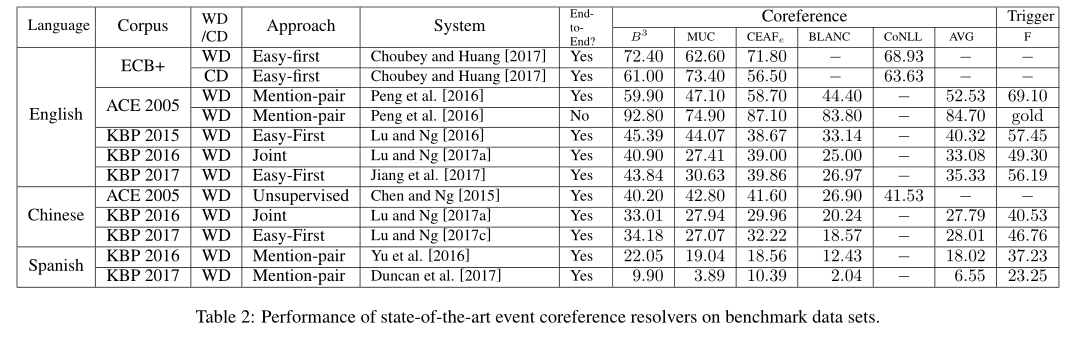
由于端到端事件共指解析器对事件提取组件提取的事件提及进行操作，事件共指性能受到事件提及(即触发词)检测性能的显著影响。考虑到触发检测的难度，一些研究人员选择在**非端到端**事件关联任务上工作，假设存在一个oracle，提供一个文档的黄金事件提及，并应用他们的事件关联算法。使用黄金提及产生的结果要比使用系统提及(即自动提取)好得多。一些研究人员认为非端到端的共指评估是不现实的，事件提及提取是**端到端**事件共指解析器的一个组成部分。

**评估指标**

最初为实体共引用评估开发的四个指标通常用于评估事件共引用解析器，即：

1. 基于链接的MUC指标；
2. 基于提及的B3指标；
3. 基于实体的CEAFe指标；
4. 基于兰德指数的BLANC指标
5. CoNLL分数，前三个指标产生的F分数的未加权平均值
6. AVG分数，前四个指标产生的F分数的未加权平均值。

**最新技术的评估**



**可能的未来工作方向**

1. 进一步研究联合模型；
2. 使用复杂的特征来改进最先进的解析器；
3. 检查是否存在特定语言的问题，这些问题可能会影响无监督、半监督和注释投影方法在涉及较少学习的语言的事件共指解析中的有效应用。此外，如果目标语言不存在大量的词汇知识库，研究获取语义知识的替代方法将是重要的。
4. 其他类型的事件共指，如两种类型的部分事件共指关系：子事件关系和成员关系。