21. 共指消解

> and even Stigand, the patriotic archbishop of Canterbury, found it advisable–"'
> 'Found WHAT?' said the Duck.
> 'Found IT,' the Mouse replied rather crossly: 'of course you know what "it"means.'
> 'I know what "it"means well enough, when I find a thing,' said the Duck: 'it's generally a frog or a worm. The question is, what did the archbishop find?'
>
> Lewis Carroll, Alice in Wonderland

语言理解的一个重要成分是知道在一篇文章中谈论的是谁。考虑以下段落:

(21.1) Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

这篇文章中每一个下划线的短语都是作者用来指一个名叫维多利亚·陈的人的。我们把像 *her* 或 *Victoria Chen* 这样的语言表达称为**提及(mention)**或**指示语(referring expression)**，被提及的话语实体(Victoria Chen)称为**所指对象(referent)**。(为了区分指示语及其指代对象，我们将前者斜体化)[1]。两个或两个以上的指示语用于指代同一话语实体，我们称之为**共指(corefer)**;因此，*Victoria Chen* 和 *she* 在(21.1)是共指。

共指是自然语言理解的重要成分。如果一个对话系统告诉用户"*联合航空下午 2 点有一班航班，国泰航空下午 4 点有一班*"，用户就会知道他说的是"*我要乘第二班*"。一个使用维基百科(Wikipedia)来回答关于居里夫人(Marie Curie)的问题的问答系统，必须在"她出生在华沙"(she was born in Warsaw)这句话中知道*她(she)*是谁。从像西班牙语这样的语言翻译而来的机器翻译系统(可以在其中插入代词)必须使用前一句子的共指来确定西班牙语句子*'" Me encanta el conocimiento" dice.'* 是否应翻译为 *'"I love knowledge", he says'*或*'"I love knowledge", she says'*。的确，此示例来自 *El País* 上的一篇有关一位女教授的实际新闻，由于共指消解不正确而在机器翻译中被误译为"he"(Schiebinger，2013)。

自然语言理解系统(和人类)根据话语模型来解释语言表达(Karttunen，1969)。**话语模型(discourse model)(**图 21.1)是一种心智模型，理解者在解释文本时会逐步构建，其中包含文本中所指实体的表示形式，实体的属性以及它们之间的关系。当首先在话语中提及一个所指对象时，我们说它的表示形式被**唤起(evoked)**进入模型中。在随后提及时，可以从模型**访问(accessed)**此表示。

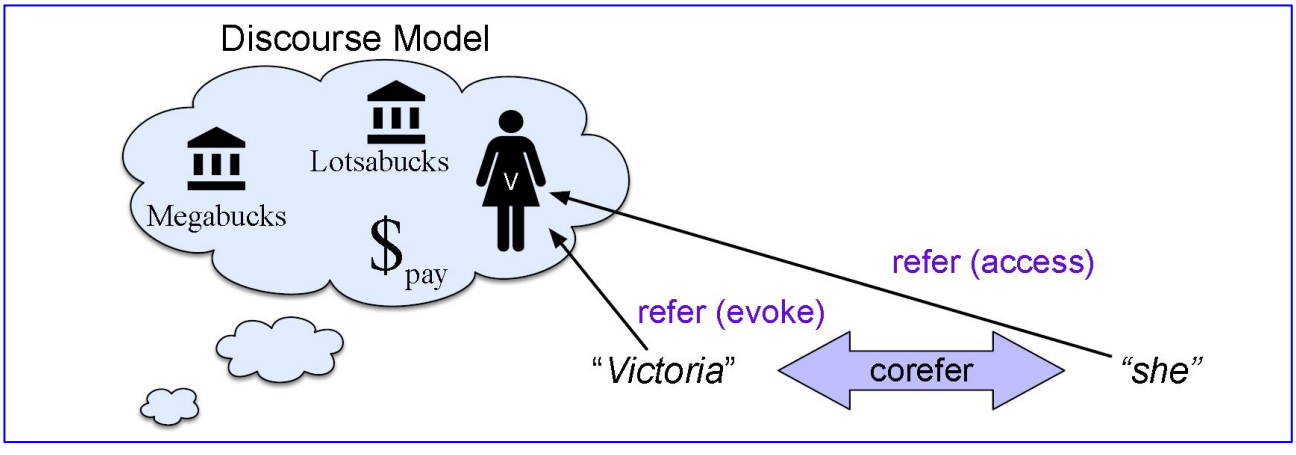

图 21-1：在话语模型中，提及如何唤起和访问话语实体

[1] 为了方便起见，有时我们会说一个指示语来指代一个所指对象，例如说 she 指代 Victoria Chen。但是，读者应该记住，我们真正的意思是，说话者是在通过说 she 来指代 Victoria Chen。

在文本中对先前已引入到话语中的实体的引用被称为**回指(anaphora)**，并且所使用的指示语被称为是**回指语(anaphor)**或照应词(anaphoric)[2]。因此，在段落(21.1)中， 代词 *she* 和 *her* 以及有定名词短语 *the 38-year-old* 都是回指语。回指语共指先前的提及(在本例中为 *Victoria Chen*)，这个提及被称为**先行词(antecedent)**。并非每个指示语都是一个先行词。在文本中只有一个提及的实体(例如(21.1)中的Lotsabucks)称为**单例(singleton)**。

在这一章中，我们将重点讨论共指消解(coreference resolution)的任务。共指消解是确定两个提及是否共指的任务，我们的意思是它们指的是话语模型中的同一实体(同一话语实体)。指示语的集合通常称为**共指链(coreference chain)**或**聚类(cluster)**。例如，在(21.1)的处理过程中，一个共指消解算法需要找到至少四个共指链，对应于图 21.1 中的话语模型中的四个实体。

1. *{Victoria Chen , her, the 38-year-old, She}*
2. *{Megabucks Banking, the company, Megabucks}*
3. *{her pay}*
4. *{Lotsabucks}*

注意，提及可以嵌套;例如，在语法上，提及 her 在语法上是另一个提及 her pay 的一部分，指的是一个完全不同的话语实体。因此，共指消解包括两个任务(尽管它们通常是共同执行的):(1)识别提及，(2)将它们聚类为共指链/话语实体。

我们说过，如果两个提及与同一个话语实体相关联，那么这两个提及就很重要。但通常我们想走得更远，确定哪个真实世界实体与此话语实体相关联。例如，提及华盛顿可能指的是美国州、首都或乔治华盛顿。对于这些句子的解释当然会有很大的不同。**实体连接(entity linking)**(Ji 和 Grishman，2011)或实体消解的任务是将话语实体映射到某些现实世界的个体任务[3]。我们通常通过映射到本体来操作实体连接或消解：世界上的实体列表，例如 gazeteer(第 15 章)。也许用于此任务的最常见的本体是 Wikipedia。每个 Wikipedia 页面都充当特定实体的唯一 ID。因此，维基化的实体连接任务(Mihalcea 和 Csomai，2007)是决定哪个维基百科页面对应的个体被提及。哪一个对应于个体的维基百科页面通过一个提及被共指。但是实体连接可以用任何本体来完成。例如，如果我们有一个基因本体，则可以将文本中提及的基因连接到本体中已消除歧义的基因名称。

在下一节中，我们将更详细地介绍共指消解的任务，并提供各种消解架构，从简单的确定性基线算法到最先进的神经模型。

然而，在讨论算法之前，我们要提及一些重要的任务，我们只会在本章的最后简要地介绍一下。首先是著名的 Winograd 模式问题(之所以这样称呼是因为它们是由 Terry Winograd 在他的论文中首先指出的)。这些实体共指消解问题被设计成很难用我们在本章中描述的解析方法来解决，而它们所需要的现实世界知识使它们成为一种具有挑战性的自然语言理解任务。例如，在下面的例子中，我们来看看如何确定代词 they 的正确先行词:

(21.2) The city council denied the demonstrators a permit because
a. they feared violence.
b. they advocated violence.

要确定代词的正确先行词，他们需要理解第二个从句是对第一个从句的解释，而且市议会可能比游行者更害怕暴力，游行者可能更倾向于提倡暴力。解决 Winograd 模式问题需要找到表示或发现必要的现实世界知识的方法。

我们在本章不讨论的一个问题是**事件共指(event coreference)**的相关任务，即确定两个事件提及(例如这两个来自 ECB+语料库的句子中的 buy 和 acquisition)是否指的是同一事件:

---

[2] 我们将按照常用的回指词的 NLP 用法，来表示有先行词的任何提及，而不是更狭义的用法来仅表示其解释取决于先行词（在更狭义的解释下，重复的名称不是回指词)的提及（如代词)。

[3] 因此，计算语言学/自然语言处理在使用“所指对象(reference)”一词时有别于形式语义领域，该领域使用“所指对象(reference)”和“共指(coreference)”两词来描述提及与现实实体之间的关系。相比之下，我们遵循功能语言学的传统，即提及指的是话语实体(Webber, 1978)，话语实体与现实世界个体之间的关系需要额外的连接步骤。

(21.3) AMD agreed to [buy] Markham, Ontario-based ATI for around $5.4 billion in cash and stock, the companies announced Monday.

(21.4) The [acquisition] would turn AMD into one of the world's largest providers of graphics chips.

事件提及比实体提及更难检测，因为它们既可以是动词性表达也可以是名词性表达。一旦检测到，用于实体的相同提及配对和提及排名模型通常会应用于事件。

一种更为复杂的共指是**话语指称(discourse deixis)**(Webber, 1988)，在其中一个回指词指回到一个话语片段，这种语段很难界定或分类，如(21.5)中的例子改编自 Webber (1991):

(21.5) According to Soleil, Beau just opened a restaurant

a. But *that* turned out to be a lie.

b. But *that* was false.

c. *That* struck me as a funny way to describe the situation.

所指对象 that 是(21.5a)中的言语行为(见第 24 章)，(21.5b)中的命题，(21.5c)中的描述方式。在本章中，我们不会为这些困难类型的非名词类先行词给出算法，但请参阅 Kolhatkar 等人(2018)的一项概述。

21.1. 共指现象：语言背景

我们现在提供一些有关共指现象的语言学背景。我们介绍了四种类型的指示语：有定和不定(definite and indefinite)的 NP、代词和名称(pronouns and names)，描述了它们是如何在话语模型中唤起和访问实体的，并讨论了回指语/先行词的语言特征(如数/性别一致，或动词语义的属性)。

21.1.1. 指示语的类型

不定名词短语(Indefinite Noun Phrases)：英语中最常见的不定指代形式是用限定词 a(或 an)来标记的，但它也可以用量词 some 甚至是限定词 this 来标记。不定指代通常引入对听众来说是新颖的话语上下文实体。

(21.6) a. Mrs. Martin was so very kind as to send Mrs. Goddard *a beautiful goose*.

b. He had gone round one day to bring her *some walnuts*.

c. I saw *this beautiful cauliflower* today.

有定名词短语(Definite Noun Phrases)：明确的指代，例如通过使用英语冠词 the 的 NP 进行引用，指的是对于听众来说是可识别的实体。听众可以识别一个实体，因为该实体先前已在文本中提及，因此已经在话语模型中表示出来：

(21.7) It concerns a white stallion which I have sold to an officer.

But the pedigree of *the white stallion* was not fully established.

或者，一个实体可以被识别，因为它包含在听众关于世界的信念集合中，或者该对象的唯一性由描述本身暗示，在这种情况下，它唤起了对象在话语模型中的表示，例如 在(21.9)中：

(21.8) I read about it in the *New York Times*.

(21.9) Have you seen the car keys?

这些最后的用法很常见。新闻通讯文本中超过一半的有定 NP 为非回指的，通常是因为它们是第一次提及的实体(Poesio 和 Vieira 1998，Bean 和 Riloff 1999)。

代词(pronouns):另一形式的有定指代是名词化，用于话语中非常突出的实体(我们将在下面讨论):

(21.10) Emma smiled and chatted as cheerfully as *she* could.

代词也可以参与**后指(cataphora)**，如(21.11)，代词在所指对象之前被提及。

(21.11) Even before *she* saw *it*, Dorothy had been thinking about the Emerald City every day.

其中，代词 she 和 it 都出现在它们的所指对象被引入之前。

代词也出现在被认为是受**绑定(bound)**的量化上下文中，如(21.12)。

(21.12) Every dancer brought *her* left arm forward.

在相关阅读材料下，she 并不是指上下文中的某个女人，而是像一个变量绑定在 every dancer 的量化

表达上。在本章中，我们不讨论代词的绑定解释。

在一些语言中，代词可以作为附在单词上的附注出现，例如下面这个西班牙语例子中的 lo(“it”)，来自 AnCora (Recasens 和 Martı,2010):

(21.13) La intención es reconocer el gran prestigio que tiene la maratón y unir**lo** con esta gran carrera.

“The aim is to recognize the great prestige that the Marathon has and join|**it** with this great race.”

指示代词(Demonstrative Pronouns):指示代词 this 和 that 既可以单独出现，也可以作为限定词出现，例如，this ingredient, that spice:

(21.14) I just bought a copy of Thoreau’s Walden. I had bought one five years ago.
That one had been very tattered; this one was in much better condition.

注意，其中 this 是有歧义的;在口语英语中，它可以是不定的，如(21.6)，也可以是有定的，如(21.14)。

零回指(zero anaphor)：在某些语言(包括中文、日语和意大利语)中，不使用代词，而是可能使用完全没有词汇实现的回指，称为零指称或零代词，如以下意大利语和日语的示例 Poesio 等人(2016)：

(21.15) EN [John]$_i$ went to visit some friends. On the way [he]$_i$ bought some wine.
IT [Giovanni]$_i$ ando a far visita a degli amici. Per via φ_i compr`o del vino.
JA [John]$_i$-wa yujin-o houmon-sita. Tochu-de φ_i wain-o ka-tta.

或者这个中文例子：

(21.16) [我] 前一会精神上太紧张.....[0] 现在比较平静了
[I] was too nervous a while ago..... [0] am now calmer.

在这些语言中，零回指使提及检测的任务复杂化。

名称(Names)：名称(如人、地点、组织等)可以用来指代话语中的新旧实体:

(21.17) a. **Miss Woodhouse** certainly had not done him justice.
b. **International Business Machines** sought patent compensation from Amazon;
IBM had previously sued other companies.

21.1.2. 信息状态

指示语的下列两种使用方式之一，称为他们的**信息状态(information status)**或信息结构。该方式是：唤起新的所指对象进入话语(引入新的信息)；或者访问旧的来自模型实体(旧信息)实体可以是**新的话语(discourse-new)**，或者**旧的话语(discourse-old)**，实际上，通常在信息上区分至少三种实体是很普遍的(Prince，1981a)：

新的 NP：
全新的 NP：这些引入的实体是新的话语和新的听众
(如 a fruit or some walnuts.)
未用的 NP：这些引入的实体是新的话语和旧的听众
(如 Hong Kong, Marie Curie, or the New York Times.)
旧的 NP：也称唤起的 NP，它们引入话语模型中已经存在的实体，因此既是旧的话语又是旧的听众的
(如 I went to a new restaurant. It was...中的 it .)

推理对象(inferrables)：这些引入的实体既不是旧的听众也不是旧的话语，但是听众可以通过基于话语中其他实体的推理来推断它们的存在。考虑以下例子:

(21.18) I went to a superb restaurant yesterday. The chef had just opened it.

(21.19) Mix flour, butter and water. Knead the dough until shiny.

不管是厨师(the chef)还是面团(the dough)都不在上述两个例子中的任何一个示例的第一句话的话语模型中，但是读者可以根据世界的观点做出**桥接推理(bridging inference)**，即这些实体应该添加到话语模型中并与餐厅和食材相关联，这个推理是基于世界知识，即餐厅有厨师和生面团是面粉和液体混合的结果(Haviland 和 Clark 1974，Webber 和 Baldwin 1992，Nissim 等人 2004，Hou 等人 2018)。

NP 的形式为其信息状态提供了强有力的线索。我们经常谈论的是实体在**给定-新的(given-new)**维度

上的位置，即所指对象是给定(given)的程度(话语中的显著性，听者更容易想到，听者可以预测)，对应于新的(new)的程度(话语中的非显著性，不可预测性)(Chafe 1976，Prince 1981b，Gundel 等人 1993)。如果一个所指对象在听者的脑海中非常突出或易于唤起人们的目光，则称其为**易于访问的(accessible)**(Ariel，2001 年)所指对象。这样的所指对象可以使用较少语言的材料来指代。例如，在话语模性中，仅当所指对象具有高度激活程度或**显著性(salience)**时，代词才被使用[4]。相比之下，不太显著的实体，如一个新的所指对象被引入话语中，将需要引入一个更长的和更明确的指示语，以帮助听众恢复所指对象。

因此，当实体首次被引入话语时，其提及的内容可能具有全名、头衔或角色、或者贴切的或限制性的相对从句，如同在(21.1)中介绍我们的主角：Megabucks Banking 的 CFO Victoria Chen。在通过话语讨论实体时，它对听众而言变得更加显著，并且平均而言，它的提及通常会变得简短且信息量较少，例如，使用简短的名字(例如 Ms. Chen)，明确的描述(the 38-year-old)或代词(she 或 her)(Hawkins 1978)。但是，这种长度的变化不是单调的，并且对话语结构很敏感(Grosz 1977b，Reichman 1985，Fox 1993)。

21.1.3. 复杂化：非指示语

许多名词短语或其他名词并非指示语，尽管它们可能有令人困惑的浅表相似性。例如，在一些最早的指代消解计算工作中，Karttunen(1969)指出，以下例子中的名词短语 a car 并不产生话语所指对象:

(21.20) Janet doesn't have a car.

并且不能通过回指词 it 或 the car 被回指：

(21.21) *It is a Toyota.

(21.22) *The car is red.

我们在此总结了四种常见的结构类型，它们不被算作共指任务中的提及，从而使提及检测任务复杂化:

同位语(Appositives):同位语结构是一个名词短语出现在一个中心名词短语的旁边，用来描述中心词。在英语中，它们经常以逗号形式出现，比如"a unit of UAL"与 NP United 相对立，或者"CFO of Megabucks Banking"与 Victoria Chen 相对立。

(21.23) Victoria Chen, CFO of Megabucks Banking, saw ...

(21.24) United, a unit of UAL, matched the fares.

同位语 NP 并不是指示语，而是充当对中心词 NP 的一种补充插入描述的功能。尽管如此，有时将这些短语与它们所描述的实体联系起来是有用的，所以像 OntoNotes 这样的数据集标记了对应关系。

表语和前名词 NP：表语或定语 NP 描述中心(head)名词的特性。在 United is a unit of UAL 中，a unit of UAL 这个 NP 描述的是 United 的属性，而不是指不同的实体。因此，它们在共指任务中没有被标记为提及。在我们的示例中，$2.3 million 和 the company's president 者两个 NP 是定语，描述了 her pay 和 the 38-year-old 的特性；范例(21.27)显示了一个中文范例，其中未提及谓词 NP(中国最大的城市; China's biggest city)。

(21.25) her pay jumped to *$2.3 million*

(21.26) the 38-year-old became *the company's president*

(21.27)上海是[中国最大的城市] [Shanghai is *China's biggest city*]

咒骂语(Expletives)：英语中的代词和其他语言中的相应代词的许多用法都不具有指称性。

这样的咒骂的或**赘述的(pleonastic)**情况包括 it is raining，在成语中如同 hit it off，或者在特殊的句法情况下如同**分裂(clefts)**(21.28a)或**外置(extraposition)**(21.28b)：

(21.28) a. It was Emma Goldman who founded Mother Earth

b. It surprised me that there was a herring hanging on her wall

类属指代(Generics):另一种表达方式是类属的指代，该表达方式不会回指到在文本中显式唤起的实体。考虑(21.29)。

(21.29) I love mangos. They are very tasty.

[4] 代词通常(但不总是)指的是在进行中的话语中一两句以内引入的实体，而有定名词短语通常指的是更早之前的实体。

这里，they 不是指代一个特定的芒果或一组芒果，而是指总体上的芒果。代词 you 也可以泛指:

(21.30) In July in San Francisco you have to wear a jacket.

21.1.4. 共指关系的语言属性

既然我们已经看到了个体指示语的语言属性，我们就转向先行词/回指词偶对的属性。理解这些属性对设计新特征和进行错误分析都有帮助。

数的一致性(Number Agreement): 指示语与它们的所指对象在数上必须大体一致;英语 she/her/he/him/his/it 为单数形式，we/us/they/them 为复数形式，而且 you 为未指定的数。因此，像 the chefs 这样的复数先行词通常不能与像 she 这样的单数回指词共指。然而，算法不能过于严格地强制数的一致性。首先，语义上的复数实体可以由 it 或 they 指代:

(21.31) IBM announced a new machine translation product yesterday.
They have been working on it for 20 years.

其次，**单数他们(singular they)**已变得越来越普遍，在单数 they 中，they 用来描述单个的个体，由于 they 对性别的中立性而常常有用。尽管最近有所增加，但 singular they 形式已经很古老了，许多世纪以来都是英语的一部分[5]。

人称一致性(Person Agreement):英语区分第一人称、第二人称和第三人称，代词的先行词必须与人称代词一致。因此第三人称代词(he, she, they, him, her, them, his, her, their)必须有第三人称先行词(以上任何一个或任何其他名词短语)。然而，像引号这样的现象会导致异常;在这个例子中，I, my 和 she 是共指的:

(21.32) “I voted for Nader because he was most aligned with my values,” she said.

性别或名词类的一致性(Gender or Noun Class Agreement):在许多语言中，所有名词都有语法性别或名词类[6]，代词通常与其先行词的语法性别一致。在英语中，这种用法只出现在第三人称单数代词上，用以区分男性(he, him, his)、女性(she, her)和非人类语法性别(it)。像 ze 或 hir 这样的非二元代词也可能出现在更近期的文本中。要知道在文本中把哪个性别与一个名字联系起来是很复杂的，可能需要对个体世界的了解。一些例子:

(21.33) Maryam has a theorem. She is exciting. (she=Maryam, not the theorem)

(21.34) Maryam has a theorem. It is exciting. (it=the theorem, not Maryam)

约束理论限制(Binding Theory Constraints)：约束理论是对同一句子中提及和先行词之间的关系进行句法约束的名称(Chomsky，1981)。稍微简化一点，像 himself 和 herself 这样的反身代词指代包含它们的最直接从句的主语(21.35)，而非反身代词不能指代这个主语(21.36)。

(21.35) Janet bought herself a bottle of fish sauce. [herself=Janet]

(21.36) Janet bought her a bottle of fish sauce. [her=Janet]

新近性(Recency):在最近的语段中引入的实体往往比从更早更远(further back)的语段中引入的实体更为显著。因此，在(21.37)中，代词 it 更可能指的是 Jim 's map 而不是 doctor 's map。

(21.37) The doctor found an old map in the captain’s chest.
Jim found an even older map hidden on the shelf. It described an island.

语法作用(Grammatical Role)：在主语位置提到的实体比在宾语位置提到的实体更显著，而在宾语位置提到的实体则比在倾斜(oblique)位置提到的实体更重要。因此，尽管(21.38)和(21.39)中的第一句话表达了大致相同的命题内容，但该代词 he 的首选指称对象随主语而有所不同-在(21.38)中是 John，而在(21.39)中是 Bill。

---

[5] 这里有一个来自莎士比亚的《错误的喜剧》的代词例子:There’ s not a man I meet but doth salute me As if I were their well-acquainted friend（我遇到的男人中，没有一个不向我致敬，就好像我是他们的知己一样)。

[6] “性别”这个词一般只用于有 2 或 3 个名词类的语言，如大多数印欧语言;许多语言，如班图语或汉语，都有大量的名词类。

(21.38) Billy Bones went to the bar with Jim Hawkins. He called for a glass of rum. [he = Billy]
(21.39) Jim Hawkins went to the bar with Billy Bones. He called for a glass of rum. [he = Jim]

动词语义学(Verb Semantics)：一些动词在语义上强调其论元之一，偏向于对后续代词的解释。比较(21.40)和(21.41)。
(21.40) John telephoned Bill. He lost the laptop.
(21.41) John criticized Bill. He lost the laptop.

这些示例仅在第一句中使用的动词有所不同，但(21.40)中的“He”通常被解析为 John，而(21.41)中的“He”被解析为 Bill。这可能是由于隐式因果关系与显著性之间的联系所致：“criticizing” 事件的隐式原因是其宾语，而“telephoning”事件的隐式原因是其主语。在这样的动词中，作为隐含原因的实体更为突出。

选择限制(Selectional Restrictions)：许多其他种类的语义知识也可以在指称偏好中发挥作用。例如，动词对其论元施加的选择限制(第 10 章)可以帮助消除所指对象，如(21.42)中所示。
(21.42) I ate the soup in my new bowl after cooking it for hours

对 it 来说有两个可能的所指对象，soup 和 bowl。然而，动词 eat 要求它的直接宾语是指可食用的东西，这一限制可能排除了 bowl 作为所指对象的可能性。

21.2. 共指任务和数据集

我们可以将共指消解的任务表述为:给定一个文本 T，找出所有实体及其之间的共指链。我们通过比较系统创建的连接与 t 上人工创建的黄金共指注释中的连接来评估我们的任务。让我们回到我们的共指示例，现在为每个共指链(聚类)使用上标数字，为聚类中的单个提及使用下标字母:

(21.43) [Victoria Chen]^1_a, CFO of [Megabucks Banking]^2_a, saw [[her]^1_b pay]^3_a jump to $2.3 million, as [the 38-year-old]^1_c also became [[the company]^2_b 's president. It is widely known that [she]^1_d came to [Megabucks]^2_c from rival [Lotsabucks]^4_a .

假设例子(21.43)是这篇文章的全部内容，那么 her pay 和 Lotsabucks 链都是单独提到的:

1. { *Victoria Chen, her, the 38-year-old, She* }
2. { *Megabucks Banking, the company, Megabucks* }
3. { *her pay* }
4. { *Lotsabucks* }

对于大多数共指评估活动，系统的输入是文章的原始文本，系统必须检测提及，然后将它们连接到聚类中。解决这个任务需要处理代词的回指(找出 her 指代 Victoria Chen),过滤掉**无所指(non-referential)**代词诸如赘述的(pleonastic)It has been ten years 中的 it),处理有定名词短语来找出 the 38-year-old 是指代 Victoria Chen,而且 the company 与 Megabucks 是一样的。我们需要处理名字，意识到 Megabucks 与 Megabucks Banking 是一样的。

确切地说，什么才算是提及？任务之间的注释链以及数据集之间的注释连接有何不同?例如，某些共指数据集不标签单例，从而使任务更加简单。解析器可以在没有单例的情况下在语料库上获得更高的分数，因为单例构成了运行文本中提到最多的内容，并且它们通常很难与无所指 NP 区别开来。某些任务使用了金牌提及检测(即为系统提供了人工标签的提及边界，而任务只是将这些金牌提及进行聚类)，这消除了从运行文本中检测和分割提及的需要。

共指的评估通常采用 CoNLL F1 分数，它结合了三个指标:MUC、B3 和 CEAFe;第 21.7 节给出了细节。

我们来谈谈一种流行的共指数据集 OntoNotes(Pradhan 等人 2007； Pradhan 等人 2007a)以及基于该数据集的 CoNLL 2012 共享任务的一些特征(Pradhan 等人 2012a)。OntoNotes 包含手动注释的中文和英文共指数据集，每个数据集大约一百万个单词，包括新闻通讯，杂志文章，广播新闻，广播对话，Web 数据和对话语音数据，以及大约 30 万个单词的阿拉伯语新闻通讯。OntoNotes 最重要的区别特征是它不标签单例，简化了共指任务，因为单例占所有实体的 60％-70％。在其他方面，它与其他共指数据集相似。

指示语 NP 被标记为提及，但类属指代和赘述代词未被标记。同位语从句不标记为单独的提及，但包含在提及中。因此，在 NP“Richard Godown, president of the Industrial Biotechnology Association”中，提及就是这个完整的短语。只有前名词修饰语是专有名词时，它们才被注释为单独的实体。因此，wheat 不是 wheat fields 中的实体，但 UN 是 UN policy 中的实体(但不是形容词诸如 American 在 American policy 中)。许多语料库都标记了更丰富的话语现象。ISNotes 语料库注释了一部分 OntoNotes 的信息状态，包括桥接示例(Hou 等人，2018)。LitBank 共指语料库(Bamman 等人 2020)包含来自 100 种不同文学小说的 210,532 个符记的共指注释，包括单例以及量化和否定的名词短语。AnCora-CO 共指语料库(Recasens 和 Martí，2010)包含 400,000 个单词，分别来自 Spanish 语(AnCora-CO-Es)和 Catalan 语(AnCora-CO-Ca)新闻数据，并在两种语言中都包含诸如话语指示等复杂现象的标签。ARRAU 语料库(Uryupina 等人，2020)包含 350,000 个英语单词，标记所有 NP，这意味着可以使用单例聚类。ARRAU 包括对话(TRAINS 数据)和小说(Pear Stories)等多种类型，并具有桥接所指对象的标签、话语指示、类属指代和模棱两可的回指关系的标签。

21.3. 提及检测

共指的第一阶段是**提及检测(mention detection)**：找到构成每个提及的文本跨度。提及检测算法通常在提议候选者提及时非常自由(即强调回忆)，并且仅在以后进行过滤。例如，许多系统在文本上运行解析器和命名实体标记器，并提取每个跨度，这些跨度可以是 NP、物主代词或命名实体。

在(21.44)中重复的示例文本中这样做：

(21.44) Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old also became the company’s president. It is widely known that she came to Megabucks from rival Lotsabucks.

可能会导致以下 13 个潜在提及列表:

Victoria Chen	$2.3 million	she
CFO of Megabucks Banking	the 38-year-old	Megabucks
Megabucks Banking	the company	Lotsabucks
her	the company’s president	
her pay	It	

最近提到的检测系统更加度量宽大。我们将在第 21.6 节中描述的基于跨度的算法首先从字面上提取所有 N-gram 跨度为 N = 10 的单词。当然，从 21.1.3 节中回想起，许多 NP(以及绝大多数随机 N-gram 跨度)都不是指示语。因此，所有这些提及的检测系统最终都需要过滤出诸如上述 it 之类的赘述词，CFO of Megabucks Banking Inc 之类的同义语，the company’s president 或 $2.3 million 之类的谓语名词。

某些过滤可以通过规则来完成。早期的基于规则的系统设计正则表达式来处理赘述词 it，就像下面来自 Lappin 和 Leass(1994)的规则，使用认知动词的字典(例如 believe, know, expect)来捕获赘述词 it 于“It is thought that ketchup…”之中，或者捕获情态形容词(例如 necessary, possible, certain, important)于“It is likely that I…”之中。这些规则有时被用作现代系统的一部分:

It is Modaladjective that S
It is Modaladjective (for NP) to VP
It is Cogv-ed that S
It seems/appears/means/follows (that) S

提及检测规则有时是专门为特定评估活动(ampaigns)设计的。例如，对于 OntoNotes，提及并没有嵌入较大的提及中，并且尽管对数字数量进行了注释，但它们很少具有核心意义。因此，对于像 CoNLL 2012(Pradhan 等人，2012a)这样的 OntoNotes 任务，一种常见的首关(first pass)基于规则的提及检测算法(Lee 等人，2013)是：

1. Take all NPs, possessive pronouns, and named entities.
2. Remove numeric quantities (100 dollars, 8%), mentions embedded in larger mentions, adjectival forms of nations, and stop words (like *there*).
3. Remove pleonastic it based on regular expression patterns.

但是，基于规则的系统通常不足以处理提及检测，因此现代系统结合了某种习得的提及检测组件，例如指称性(referentiality)分类器，回指性(anaphoricity)分类器(检测 NP 是否是一个回指词)或全新的话语(discourse-new)分类器-检测提及是否是全新的话语以及将来的回指的潜在先行词。

例如，一个**回指性检测器(anaphoricity detector)**可以从任何在手工标签的数据集(如 OntoNotes、ARRAU 或 AnCora)中标签为回指指示语的跨度中提取积极的训练示例。任何其他 NP 或命名实体都可以标记为一个消极训练示例。回指性分类器利用候选提及的中心词、周边词、确定性、活动性、长度、在句子/话语中的位置等特征，其中许多特征在 Ng 和 Cardie (2002a)的早期著作中首次提出;有关特征的更多信息，请参阅 21.5 节。

指称性或回指性检测器可以作为过滤器运行，其中只有被归类为回指性或指称性的提及被传递给共指系统。在我们上面的例子中，这种过滤提及检测系统的最终结果可能是以下 9 个潜在提及的过滤集:

Victoria Chen	her pay	she
Megabucks Bank	the 38-year-old	Megabucks
her	the company	Lotsabucks

然而，事实证明，基于回指性或指称性分类器对提及内容进行硬过滤会导致性能较差。如果回指性分类器阈值设置得太高，就会过滤掉太多的提及，召回率受影响；如果分类器阈值设置得太低，就会包含太多的赘述词或非指代提及，从而精度受影响。

相反，现代方法是在单个端到端模型中联合执行提及检测、回指性和共指(Ng 2005b，Denis 和 Baldridge 2007，Rahman 和 Ng 2009)。例如，Lee 等人(2017)(2018)系统中的提及检测是基于单个端到端神经网络的，该网络计算每个提及为指称对象的得分，计算两个提及为共指的得分，并将它们组合起来以做出决定，用单个端到端的损失就可以训练所有这些分数。我们将在第 21.6 节中详细介绍此方法。[7]

尽管有这些进步,正确检测所指对象的提及似乎仍然是一个尚未解决的问题,因为系统错误地将赘述代词(如 it)和其他非指称 NP 标记为共指，所以这个系统是现代共指消解系统的一大错误来源(Kummerfeld 和 Klein 2013，Martschat 和 Strube 2014，Martschat 和 Strube 2015，Wiseman 等人 2015，Lee 等人 2017)。

因此，提及，指称性或回指性检测是一个重要的开放研究领域。其他知识来源可能会有所帮助，尤其是与无监督和半监督算法结合使用时，也可以减少标记数据集的费用。在早期的工作中，例如 Bean 和 Riloff(1999)学习了表征回指或非回指 NP 的模式(通过提取和归纳文本中的第一个 NP 来保证它们是非回指的)。Chang 等人(2012)寻找在训练数据中频繁出现但从未作为金牌提及出现的中心词(head)名词，以帮助找到非指称 NP。Bergsma 等人(2008b)使用网络计数作为一种半监督的方式来增强英语 it 的回指性检测的标准特征，这是一项重要的任务，因为它既常见又有歧义。在 1/4 到 1/2 之间，它的例子是非回指的。请考虑以下两个示例：

(21.45) You can make [it] in advance. [anaphoric]

(21.46) You can make [it] in Hollywood. [non-anaphoric]

“make it”中的“it”是非回指的，是习语“make it”的一部分。Bergsma 等人(2008b)将每个例子周围的上下文转换成模式，如(21.45)中的“make * in advance”和(21.46)中的“make * in Hollywood”。然后他们使用谷歌 n-gram 来枚举所有可以在模式中取代 it 的单词。非回指上下文往往只把 it 放在通配符位置上，而回指上下文出现在许多其他 NP 中(例如，make them in advance 和 make it in advance 在它们的数据中出现的频率一样，但 make them in Hollywood 根本就没有出现)。这些 N-gram 上下文可以作为受监督的回指性分类器的特征。

[7] 一些系统试图完全避免提及检测或回指检测。对于像 OntoNotes 这样的不标记单例的数据集，过滤掉非指代提及的另一种方法是运行共指消解，然后简单地删除所有未被其他提及所共指的候选提及。这可能不像对指代性进行显式建模那样有效，也无法解决检测单例的问题，这对于诸如实体连接之类的任务很重要。

21.4. 共指算法的架构

用于共指的现代系统是基于监督的神经机器学习，该监督是通过手工标记的数据集(如 OntoNotes)进行监督的。在本节中，我们概述现代系统的各种体系结构，使用的方法是 Ng(2010)分类，该分类从两个方面来区分不同的算法：第一方面，他们是否以以下方式做出每个共指决定：基于实体(代表话语模型中的每个实体)或仅基于提及(独立考虑每个提及)；第二方面，他们是否使用排名模型直接比较潜在的先行词。之后，我们将在第 21.6 节中详细介绍一种最新算法。

21.4.1. 提及配对架构

我们从**提及配对(mention-pair)**体系结构开始，这是最简单、最具影响力的共指体系结构，它引入了更复杂算法的许多特性，尽管其他体系结构的性能更好。提及配对体系结构基于一个分类器——顾名思义——给出一对提及、一个候选回指词和一个候选先行词，并做出二元分类决策:是否共指(coreferring)。

考虑一下这个分类器在我们的例子中对代词 she 的任务，并假设图 21.2 中略微简化的潜在先行词集。

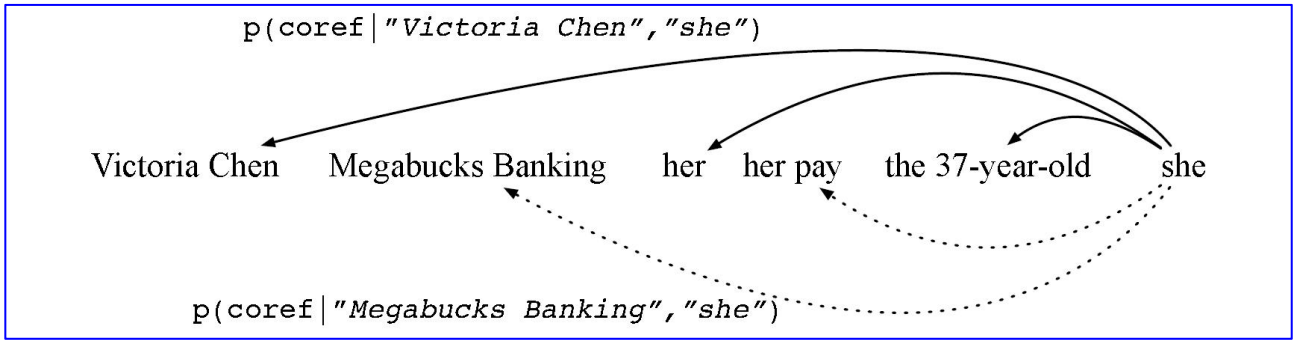

图 21-2：提及配对分类器

图注：对于每对提及(如 she)和一个潜在的先行提及(如 Victoria Chen 或 her)，提及配对分类器分配一个共指连接的概率。

对于每一个先前的提及(Victoria Chen, Megabucks Banking, her，等等)，二元分类器计算一个概率:是否提及是 she 的先行词。我们希望实际先行词(Victoria Chen, her, the 38-year-old)有较高的概率，以及非先行词(Megabucks Banking, her pay)有较低的概率。早期的分类器使用手工构建的特征(第 21.5 节);最近的分类器使用神经表示学习(第 21.6 节)。

对于训练，我们需要一个启发式的方法来选择训练样本;由于文献中的大多数提及配对都是不相关的，选择每一对都会导致大量的负样本过剩。从(Soon 等人，2001)开始，最常见的启发式方法是选择最接近的先行词作为正面例子，而选择介于两者之间的所有成对先行词作为负面例子。更正式地说，对每一个回指提及 m_i，我们创建：

- 一个正例(m_i，m_j)，其中 m_j 是与 m_i 最接近的先行词
- 一个负例(m_i，m_k)，其中每个 m_k 都在 m_j 和 m_i 之间

因此，对于回指词 she，我们会选择(she, her)作为正例而不是负例。同样，对于回指词 the company，我们选择(the company, Megabucks)作为正例，而选择(the company, she) 、(the company, the 38-year-old) 、(the company, her pay) 、和 (the company, her)作为负例。

一旦训练好分类器，就把它应用到聚类步骤中的每个测试句子中。对于文档中每个提及 i，分类器都会考虑先前 i-1 个提及中的每一个提及。在最近优先(closest-first)聚类(Soon 等人，2001)中，分类器从右到左运行(从提及 i-1 到提及 1)，而且概率> 0.5 的第一个先行词被连接到 i。如果无先行词有概率> 0.5，则对 i 来说没有先行词被选择。在最好优先(best-first)聚类方法中，分类器是运行在所有 i-1 个先行词之上的，而且最可能的优先提及被选为 i 的先行词。成对关系的可传递闭包被视为聚类。

尽管提及配对模型具有简单性的优点，但它有两个主要问题。首先，分类器不会直接对候选先行词之间进行比较，因此没有经过训练就可以在两个可能的先行词之间进行决策，判别实际上哪个更好。其次，它忽略了话语模型，只关注提及，而不关注实体。每个分类器的决策制定完全局限于在该配对中，不考虑同一实体的其他提及。接下来的两个模型分别解决了这两个缺陷之一。

21.4.2. 提及排名架构

提及排名模型直接将候选先行词相互比较，为每个回指词选择得分最高的先行词。

在早期的公式中，对于提及 i，分类器会确定先前提及{1，...，i-1}中哪个是先行词(Denis 和 Baldridge，2008)。但是，假设 i 实际上不是回指词，并且没有任何一个先行词都将不被选择？这样的模型将需要在 i 上运行单独的回指性分类器。取而代之的是，更好的方法是与以及单个损失共同学习回指检测和共指(Rahman 和 Ng，2009)。

因此，在现代提及排名系统中，对于第 i 个提及(回指词)，我们有一个关联的随机变量 yi，其范围为值 Y(i)= {1，...，i-1，ϵ }。值ϵ是一个特殊的虚拟提及，表示我没有先行词(即：要么是全新的话语和新共指链的开头，要么是一个非回指词)。

在测试时，对于给定的提及 i，模型在所有先行词(加上ϵ)上计算一个 softmax，给出每个候选先行词(或无)的概率。

图 21.3 给出了单个候选回指词 she 的计算示例。

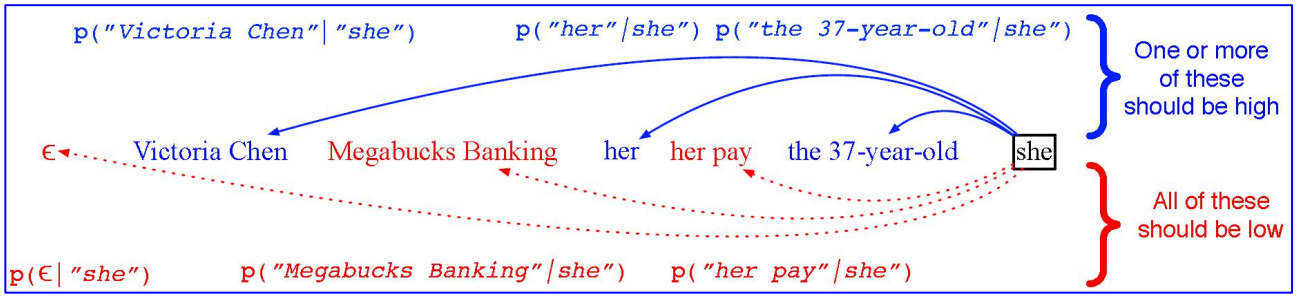

图 21-3：提及排名系统的概率分布

图注：对于每个候选的回指提及(如 she)，提及排名系统对之前所有提及加上特殊的虚拟提及ϵ分配一个概率分布。

一旦为每个回指词分类了先行词，就可以在成对的决策上运行传递闭包，以获得完整的聚类。

在提及排名模型中，训练比提及配对模型更为棘手，因为对于每个回指词，我们都不知道用于训练的所有可能金牌先行词中的哪一个。取而代之的是，每次提及的最佳先行词是潜在的。也就是说，对于每一个提及，我们都有合法的金牌先行词的一个完整聚类可供选择。早期的工作使用启发式方法来选择一个先行词，例如，选择最接近的先行词作为金牌先行词，并在两个句子的窗口中选择所有非先行词作为负例(Denis 和 Baldridge，2008)。存在多种对潜在先行词建模的方法(Fernandes 等人 2012，Chang 等人 2013，Durrett 和 Klein 2013)。最简单的方法是对任何一个先行词都给出一个信用分数(通过对它们全部加总)，其损失函数可以优化在金牌聚类中所有正确的先行词的似然度(Lee 等人，2017)。第 21.6 节有详细信息。

提及排名模型可以用手工构建的特征或神经表示学习(也可能包含一些手工构建的特征)来实现。我们将在第 21.5 节和第 21.6 节中探讨这两个方向。

21.4.3. 基于实体的模型

提及配对模型和提及排名模型都做出有关提及的决定。相比之下，基于实体的模型将每个提及不连接到先前提及，而是连接到先前的话语实体(提及的聚类)。

只需使分类器根据提及聚类而不是单个提及做出决策，就可以将提及排名模型转变为实体排名模型(Rahman 和 Ng，2009)。

对于传统的基于特征的模型，这可以通过在聚类上提取特征来完成。聚类的大小以及其“形状”都是有用的功能，它是聚类中提及类型的序列，即符记(P)合适的、(D)有定、(I)不定、(Pr)代词，所以这样一个由{Victoria, her, the 38-year-old)组成的聚类将具有 P-Pr-D 的形状(Björkelund 和 Kuhn，2014)。包含提及配对分类器的基于实体的模型可以使用提及配对概率的总计作为特征，例如计算两个聚类中所有提及配对的共指平均概率(Clark 和 Manning，2015)。

神经模型可自动学习聚类的表示，例如，通过在聚类提及序列上使用 RNN 来编码与聚类表示相对应的状态(Wiseman 等人，2016)，或者通过学习成对聚类的分布式表示，即对提及配对的习得表示进行池化。

然而，尽管基于实体的模型表达能力更强，但在实践中使用集聚类级别的信息并没有带来很大的性能提升，因此提及排名模型仍然是更常用的。

21.5. 使用手工构建特征的分类器

手工设计的功能在共指中起着重要的作用，无论是作为神经预分类器中分类的唯一输入，还是作为先进神经系统中使用的自动表示学习的增强(就像我们将在第 21.6 节中描述的那样)。在本节中，我们将介绍在逻辑回归，SVM 或随机森林分类器中用于共指解析的常用功能。

给定一个回指词提及和一个潜在的先行词提及，大多数基于特征的分类器都使用三种特征的类型：

(i)回指词的特征，

(ii)候选先行词的特征，

(iii)配对之间关系的特征。

基于实体的模型可以额外使用两个附加类：

(iv)在先行词实体聚类中的所有提及的特征，

(v)在先行词实体聚类中的回指与提及之间关系的特征。

Features of the Anaphor or Antecedent Mention		
First (last) word	Victoria/she	First or last word (or embedding) of antecedent/anaphor
Head word	Victoria/she	Head word (or head embedding) of antecedent/anaphor
Attributes	Sg-F-A-3-PER/Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of (antecedent/anaphor)
Length	2/1	length in words of (antecedent/anaphor)
Grammatical role	Sub/Sub	The grammatical role—subject, direct object, indirect object/PP—of (antecedent/anaphor)
Mention type	P/Pr	Type: (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun) of antecedent/anaphor
Features of the Antecedent Entity		
Entity shape	P-Pr-D	The 'shape' or list of types of the mentions in the antecedent entity (cluster), i.e., sequences of (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun.
Entity attributes	Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of the antecedent entity
Tntecedent cluster size	3	Number of mentions in the antecedent cluster
Features of the Pair of Mentions		
Longer anaphor	F	True of anaphor is longer than antecedent
Pairs of any features	Victoria/she, 2/1, Sub/Sub, P/Pr, etc .	For each individual feature, pair of type of antecedent+ type of anaphor
Fentence distance	1	The number of sentences between antecedent and anaphor
Mention distance	4	The number of mentions between antecedent and anaphor
i-within-I	F	Anaphor has i-within-i relation with antecedent
Cosine		Cosine between antecedent and anaphor embeddings
Appositive	F	True if the anaphor is in the syntactic apposition relation to the antecedent. Useful even if appositives aren't mentions (to know to attach the appositive to a preceding head)
Features of the Pair of Entities		
Exact String Match	F	True if the strings of any two mentions from the antecedent and anaphor clusters are identical.
Head Word Match	F	True if any mentions from antecedent cluster has same headword as any mention in anaphor cluster
Word Inclusion	F	All words in anaphor cluster included in antecedent cluster
Features of the Document		
Genre/source	N	The document genre— (D)ialog, (N)ews, etc,

图 21-4：基于特征的共指算法的一些常见特征

图注：图中包括回指语“she”和潜在先行词“Victoria Chen”的值。

图 21.4 显示了常用特征的选择，并显示了我们示例句中潜在的回指词“she”和潜在先行词“Victoria Chen”需要计算的值，重复如下:

(21.47) **Victoria Chen**, CFO of Megabucks Banking, saw her pay jump to $2.3 million,
as the 38-year-old also became the company’s president.
It is widely known that **she** came to Megabucks from rival Lotsabucks.

先前的工作发现特别有用的特性是精确字符串匹配、实体中心词一致、提及距离，以及(对于代词)精确属性匹配和 i 在 i 内(i-within-i)，以及(对于名词和专有名称)单词包含和余弦。对于词汇特征(如中心词)，通常只使用出现次数足够多的词(可能超过 20 次)，对于罕见的词就退回到词类。

在基于特征的系统中，使用特征的连接是至关重要的;一项实验表明，从分类器中的单个特征移动到多个特征的连接将 F1 增加 4 个点(Lee 等人，2017)。具体的连接可以手工设计(Durrett 和 Klein, 2013)，所有成对的特征都可以被连接(Bengtson 和 Roth, 2008)，或者可以使用决策树或随机森林分类器学习特征连接(Ng 和 Cardie 2002a, Lee 等人 2017)。

最后，其中一些特征也可以用于神经模型。现代神经系统(第 21.6 节)使用上下文词嵌入，所以它们不能从添加诸如字符串或中心词匹配、语法角色或提及类型等浅层特征中受益。然而，像提及长度、提及之间的距离或类型等其他功能可以很好地补充神经上下文嵌入模型。

21.6. 一种神经提及排名算法

在本节中，我们描述 Lee 等人(2017)的神经提及排名系统。这个端到端系统并没有单独的提及检测步骤。取而代之的是，它将可能达到设定长度的所有文本跨度(即长度为 1,2,3 ... N 的所有 n-gram)视为可能提及的内容[8]。

给定一个带有 T 个单词的文档 D，该模型会考虑所有 N = T(T-1)/ 2 个文本跨度都达到某个长度(在 Lee 等人(2018)的版本中，该长度为 10)。每个跨度 i 都从单词 START(i)开始，到单词 END(i)结束。

任务是为每个跨度 i 分配一个先行词 y_i，一个随机变量其值范围为 Y(i)= {1，...，i-1，e}；每个以前的跨度和一个特殊的虚拟符记 e。选择虚拟符记意味着 i 没有先行词，这是因为 i 是全新的话语并开始了一个新的共指链，或者因为 i 是一个非回指词。

对于跨度 i 和 j 的每一对，系统都会为跨度 i 和跨度 j 之间的共指连接分配分数 s(i，j)。然后，系统学习跨度 i 的先行词分布 $P(y_i)$：

$$P(y_i)=\frac{exp(s(i,\ y_i))}{\sum_{y' \in Y(i)} exp(s(i,\ y_i)} \tag{21.48}$$

这个分数 s(i, j)包括三个因素:

m(i)：跨度 i 是否提及;

m(j)：跨度 j 是否提及;

c(i, j)：j 是否为 i 的先行词。

$$s(i, j) = m(i) + m(j) + c(i, j) \tag{21.49}$$

对于虚拟先行词 e，分数 s(i, e)固定为 0。这样，如果任何非虚拟分数是正的，则模型预测出得分最高的前先行词；如果所有分数都是负的，则它弃权。评分函数 m(i)和 c(i, j)是基于一个表示跨度 i 的向量 $\mathbf{g}_i$：

$$m(i) = w_m \cdot FFNN_m(\mathbf{g}_i) \tag{21.50}$$

$$c(i,j) = w_c \cdot FFNN_c([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \circ \mathbf{g}_j, \Phi(i,j)]) \tag{21.51}$$

先行词分数 c(i, j)作为跨度 i 和 j 的表示的输入，但也作为两个跨度相互之间的逐元素相似性 $\mathbf{g}_i \circ \mathbf{g}_j$(其中∘是逐元素乘法)。先行词分数 c 还考虑了一个特征向量 Φ(i,j)，它编码有用的特征诸如提及距离，以及关于说话者和类型的信息。

可以使用 biLSTM 或 BERT 计算每个跨度 $\mathbf{g}_i$ 的表示形式。

在 biLSTM 版本中，范围 i 的表示 $\mathbf{g}_i$ 是一个拼接，这个拼接由四项内容生成：跨度的起始符记的 biLSTM

[8] 但是由于这种潜在提及的数量使得算法非常缓慢和笨拙(模型的文档长度是 O(t4))，在实践中，各种版本的算法会找到方法来剪除可能的提及，本质上使用提及分数作为提及检测器。

隐藏表示，跨度的结束符记的 biLSTM 隐藏表示，中心词的基于注意力的表示，仅包含一个特征(跨度 i 的长度)的特征向量。

$$g_i = [\mathbf{h}_{START(i)}, \mathbf{h}_{END(i)}, \mathbf{h}_{ATT(i)}, \phi(i)] \tag{21.52}$$

这些计算如下。对于输入的每个单词 w_t，biLSTM 的输出为 h_t：

$$\overrightarrow{\mathbf{h_t}} = \text{LSTM}^{\text{forward}}(\overrightarrow{\mathbf{h_{t-1}}}, \mathbf{w_t})$$

$$\overleftarrow{\mathbf{h_t}} = \text{LSTM}^{\text{forward}}(\overleftarrow{\mathbf{h_{t+1}}}, \mathbf{w_t})$$

$$\mathbf{h_t} = [\overrightarrow{\mathbf{h_t}}, \overleftarrow{\mathbf{h_t}}] \tag{21.53}$$

注意力表示照常被计算;系统学习一个权向量 w_α，用 FFNN 变换的隐藏状态 h_t 计算其点积:

$$\alpha_t = \mathbf{w}_\alpha \cdot \text{FFNN}_\alpha(\mathbf{h_t}) \tag{21.54}$$

注意力得分通过 softmax 归一化为一个分布:

$$\alpha_{i,t} = \frac{exp(\alpha_t)}{\sum_{k=ATART(i)}^{END(i)} exp(\alpha_k)} \tag{21.55}$$

然后，利用注意力分布来创建一个向量 $\mathbf{h}_{ATT(i)}$，它是跨度 i 的单词的注意力加权总和:

$$\mathbf{h}_{ATT(i)} = \sum_{t=ATART(i)}^{END(i)} \alpha_{i,t} \cdot \mathbf{w}_t \tag{21.56}$$

图 21.5 由 Lee 等人(2017)给出了跨度表示和提及分数的 biLSTM 计算。

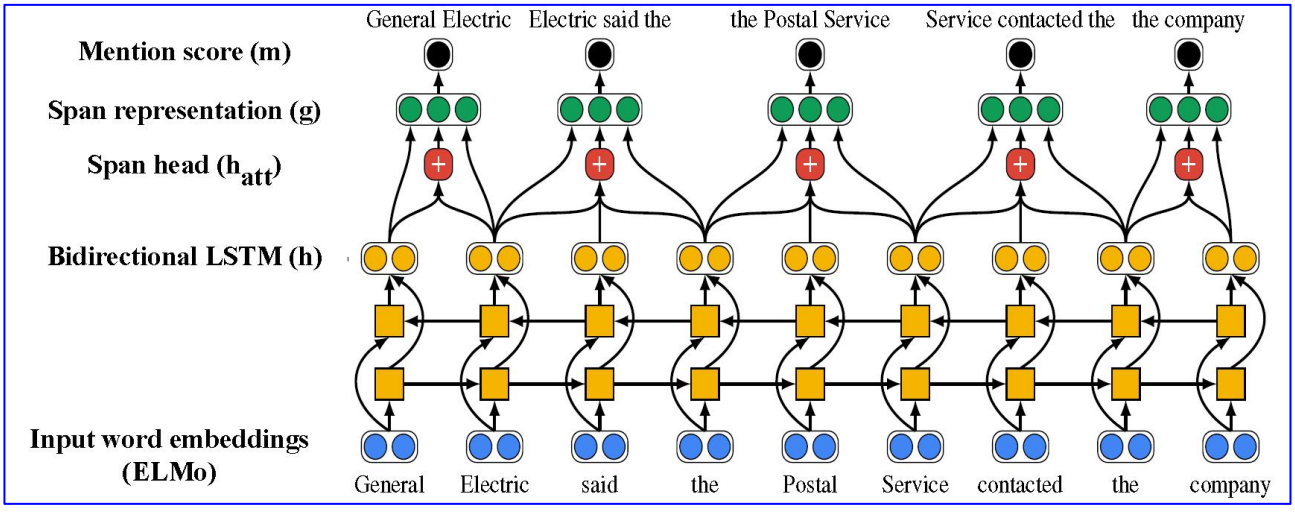

图 21-5：跨度表示和提及分数的计算

图注：计算跨度表示(使用 biLSTM 编码器)和 Lee 等人(2017)端到端共指模型中的提及分数。该模型考虑了所有跨度的最大宽度;图中显示了其中的一小部分。图形根据 Lee 等人(2017)。

在 BERT 版本中，整个 biLSTM 编码器被替换成 BERT。相反，跨度表示 g_i 的计算方法是拼接跨度的第一个和最后一个词片，加上跨度中所有词片的注意力版本(Joshi 等人，2019)。

图 21.6 给出了图 21.5 中的例句中 the company 三种可能的先行词得分 s 的计算结果。

在推断时，通常使用某种方法来修剪提及(例如，使用提及得分 m 作为过滤器，以仅保留最佳的少数提及作为函数，例如句子长度 T 的 0.4T)。然后，以正向方式计算每个文档的先行词的联合分布。最后，我们可以对先行词进行传递闭包，以便为文档创建最终的聚类。

对于训练，每次提及我们都没有一个金牌先行词。取而代之的是，共指标签仅给我们每个共指提及的整个聚类。提及有一个潜在的先行词。因此，我们使用损失函数来最大化任何合法先行词的共指概率之和。对于给定的提及 i 和可能的先行词 Y(i)，令 GOLD(i)为包含 i 的金牌聚类中的提及集合。由于提及的集合发生在 i 之前是 Y(i)，该金牌聚类中也出现在 i 之前的一组提及为 $Y(i) \cap GOLD(i)$。因此，我们希望最大化：

$$\sum_{\hat{y} \in Y(i) \cap GOLD(i)} P(\hat{y}) \tag{21.57}$$

如果提到 i 不在一个金牌聚类中，则 GOLD(i) = ε 。

为了把这个概率变成一个损失函数，我们将使用我们在第 5 章的公式 5.11 中定义的交叉熵损失函数，取其概率的-log。如果我们将所有提及的相加，我们得到训练的最终损失函数:

$$L = \sum_{i=2}^{N} -log \sum_{\hat{y} \in Y(i) \cap GOLD(i)} P(\hat{y}) \tag{21.58}$$

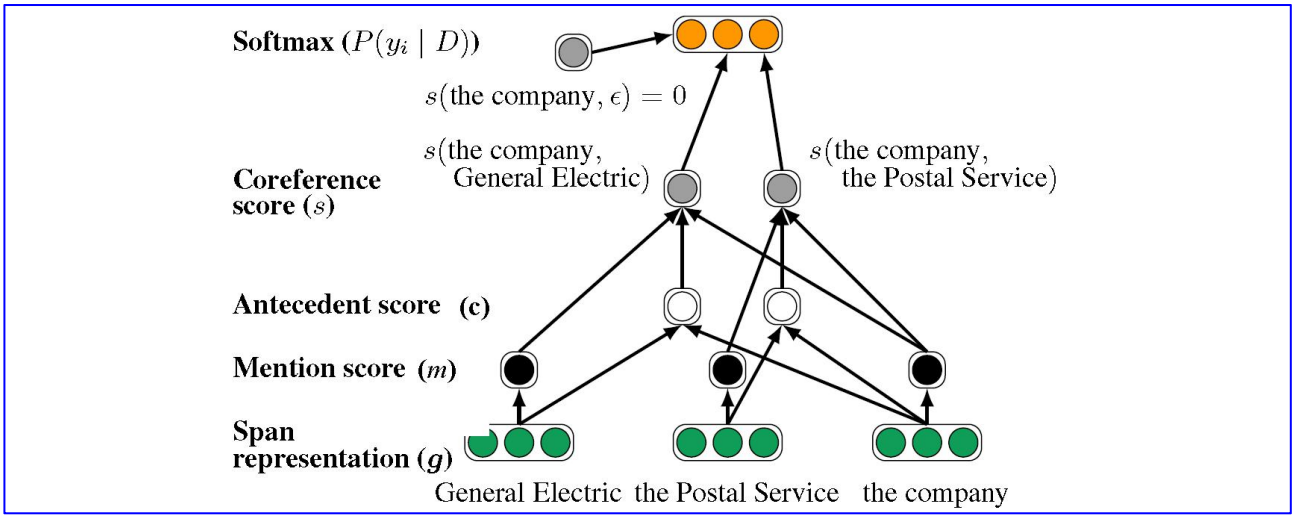

图 21-6：先行词得分的计算

图注：在图 21,5 中的例句中 the company 三种可能的先行词得分 s 的计算，图形根据 Lee 等人(2017)。

图 21.7 显示了模型的预测示例，显示了 Lee 等人(2017)发现的与传统语义中心词相关的注意力权重。请注意，该模型对第二个例子的理解是错误的，这大概是因 attendants 和 pilot 可能有附近的单词嵌入。

We are looking for (a region of central Italy bordering the Adriatic Sea). (The area) is mostly mountainous and includes Mt. Corno, the highest peak of the Apennines. (It) also includes a lot of sheep, good clean-living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them.
(The flight attendants) have until 6:00 today to ratify labor concessions. (The pilots') union and ground crew did so yesterday.

图 21-7：Lee 模型的样本预测

图注：图中每个例子一个聚类，显示一个正确的例子和一个错误的例子。粗体和圆括号的跨度是预测聚类中的提到。单词上红色的数量表示在公式(21.55)中的中心词发现(head-finding)的注意力权重 $a_{i,t}$。图改编自 Lee 等人(2017)。

21.7. 共指消解的评估

我们评估共指算法模型——理论上，将系统产生的一组假设链或聚类 H 与人类标签的一组金牌或指代链或聚类 R 进行比较，并报告精度和召回率。

然而，有各种各样的方法可以进行这种比较。事实上,有 5 个常用度量来评估共指算法：基于连接的 MUC(Vilain 等人，1995)，BLANC(Recasens 和 Hovy 2011，Luo 等人 2014)度量，基于提及的 B3 度量(Bagga 和 Baldwin,1998)，基于实体的 CEAF 度量(Luo，2005)，基于连接的实体察觉 LEA 度量(Moosavi 和 Strube，2016)。

让我们来探讨两个指标。**MUC F 度量(MUC F-measure)**(Vilain 等人，1995)基于 H 和 R 共有的共指连接(成对提及)的数量。精度是共有连接的数量除以 H 中的连接数量。召回率是共有连接数除以 R 中的连接数量；这使 MUC 偏向于产生大型链(且实体较少)的系统，并且忽略了单例，因为它们不涉及连接。

B3 是基于提及而不是基于连接的。对于指代链中的每个提及，我们都计算精度和召回率，然后对文档中所有 N 个提及项进行加权求和，以计算整个任务的精度和召回率。对于给定的提及 i，令 R 为包括 i 的指代链，H 为具有 i 的假设链。H 中正确提及的集合是 H∩R。因此，提及 i 的精度为 |H∩R| / |H|，并且

提及 i 的召回率是 |H∩R| / |R|，总的精度为提及 i 的精度的加权总和，由权重 w_i 加权。总的召回率是提及 i 的召回率的加权总和，由权重 w_i 加权。等效地：

$$\text{Precision} = \sum_{i=1}^{N} w_i \frac{\#\ of\ correct\ mentions\ in\ hypothesis\ chain\ containing\ entity_i}{\#\ of\ mentions\ in\ hypothesis\ chain\ containing\ entity_i}$$

$$\text{Recall} = \sum_{i=1}^{N} w_i \frac{\#\ of\ correct\ mentions\ in\ hypothesis\ chain\ containing\ entity_i}{\#\ of\ mentions\ in\ reference\ chain\ containing\ entity_i}$$

每个实体的权重 wi 可以设置为不同的值，以产生不同版本的算法。

根据 Denis 和 Baldridge(2009)的建议，CoNLL 共指竞赛是根据 MUC、CEAF-e 和 B3 的平均值进行评分的(Pradhan 等人 2011，Pradhan 等人 2012b)，因此在许多评估活动中，报告这 3 个指标的平均值是很常见的。请参阅 Luo 和 Pradhan(2016)了解整个指标集的详细描述;应该使用这些指代的实现，而不是试图从头开始重新实现(Pradhan 等人，2014)。

已经提出了处理特定共指领域或任务的替代度量。例如，考虑解决对命名实体(个人、组织、地缘政治实体)的提及的任务，这可能对信息提取或知识库的完成很有用。假设链正确地包含了一个指代实体的所有代词，但没有名称本身的版本，或者连接了错误的名称，对于此任务没有用。相反，我们可能想要如下两个度量之一：根据提及的信息丰富性(每个人的名字具有最多的信息)对每个提及进行加权的度量(Chen 和 Ng，2013)；仅当一个假设包含名称的至少一个变体时才认为该假设与金牌链匹配的度量(NEC F1 度量)(Agarwal 等人，2019)。

21.8. Winograd 模式问题

从该领域早期开始，研究人员就注意到，一些共指案例相当困难，似乎需要世界知识或复杂的推理才能解决。Winograd(1972)用下面的例子指出了这个问题:

(21.59) The city council denied the demonstrators a permit because
a. they feared violence.
b. they advocated violence.

Winograd 注意到，大多数读者喜欢代词 they 的先行词在(a)中是 the city council，而在(b)中是 the demonstrators。他认为，这需要理解第二个条款是对第一个条款的解释，同时我们的文化框架表明，市议会可能比示威者更害怕暴力，示威者可能更支持暴力。

为了使自然语言处理领域更多地关注涉及世界知识和常识推理的方法，Levesque(2011)提出了一个挑战任务，名为 **Winograd 模式(Winograd Schema)**挑战[9]。挑战任务中的问题是共指问题，该问题被设计成容易被人类读者消除歧义，但希望不能通过简单的技术(如选择限制，或其他基本的词关联方法)来解决。

这些问题被设定为一对陈述，它们在一个单词或短语中存在差异和一个共指问题:

(21.60) The trophy didn't fit into the suitcase because it was too **large**.
Question: What was too **large**? Answer: The trophy

(21.61) The trophy didn't fit into the suitcase because it was too **small**.
Question: What was too **small**? Answer: The suitcase

这些问题具有以下特点:

1.问题各有两个参与方

2.代词优先指代一个参与方，但是在语法上也可以指代另一参与方

3.一个问题询问代词指代的是哪一个参与方

4.如果问题中的一个单词被更改，则人类偏爱的答案将更改为另一参与方

解决问题可能需要的世界知识种类可能会有所不同。在上述的奖杯/手提箱例子中，它是有关物理世界的知识；较大的物体无法容纳较小的物体。在最初的 Winograd 句子中，它是对诸如政治人物和抗议者之类的社会角色的刻板印象。在以下示例中，它的知识是关于人类行为，例如**轮流发言(turn-taking)**或感谢。

[9] Levesque 的征集很快得到了 Levesque 等人(2012)、Rahman 和 Ng(2012)、IJCAI 会议上的一项竞赛(Davis 等人 2017)，以及所谓 WNLI 问题的自然语言推理版本(Wang 等人，2018)。

(21.62) Bill passed the gameboy to John because his turn was [over/next].
Whose turn was [over/next]? Answers: Bill/John

(21.63) Joan made sure to thank Susan for all the help she had [given/received].
Who had [given/received] help? Answers: Susan/Joan.

尽管 Winograd 模式被设计为需要常识推理，但可以通过对 Winograd 模式句子进行微调的预训练语言模型来解决原始问题集合的很大一部分(Kocijan 等人，2019)。大型的预训练语言模型可编码大量的世界常识知识！因此，目前的趋势是提出新的数据集，这些数据集具有越来越难的类似 Winograd 的共指消解问题，如 KNOWREF (Emami 等人，2019)，例如:

(21.64) Marcus is undoubtedly faster than Jarrett right now but in [his] prime the gap wasn't all that big.

最后，语言建模与知识的某种结合似乎会取得丰硕的成果;事实上，基于知识的模型似乎不太适合 Winograd 模式训练集中的词汇特质(Trichelair 等人，2018).

21.9. 共指的性别偏见

就像语言处理的其他方面一样，共指模型也表现出性别和其他偏见(Zhao 等人 2018a,Rudinger 等人 2018, Webster 等人 2018)。

例如，WinoBias 数据集(Zhao 等人，2018a)使用 Winograd Schema 范例的变体来测试共指算法在将性别代词与符合文化刻板印象的先行词联系起来时的偏向程度。正如我们在第 6 章中总结的那样，嵌入在训练测试中复制了社会偏见，例如将男性与具有历史上刻板的男性职业(例如医生)、将女性与具有陈规定型的女性职业(例如秘书)联系起来(Caliskan 等人 2017，Garg 等人 2018)。

WinoBias 句子包含两个提及，分别对应于定型为男性和定型为女性的职业，以及必须与之相关的性别代词。句子不能通过代词的性别来消除歧义，但是一个有偏见的模型可能会被这个线索分心。这是一个示例语句：

(21.65) The secretary called the physician*i* and told him*i* about a new patient [pro-stereotypical]

(21.66) The secretary called the physician*i* and told her*i* about a new patient [anti-stereotypical]

Zhao 等人(2018a)认为，如果关联代词与性别定型的职业相一致(例如，在(21.65)中的 him 和 physician)相比不一致(例如，在(21.66)中的 her 和 physician)的准确性更高，则考虑这个共指系统具有偏见性。他们表明，所有架构(基于规则的、基于特征的机器学习和端到端神经系统)的共指系统都表现出显著的偏见，在反定型的情况下平均表现差 21 个 F1 点。

这种偏见的一个可能原因是，女性实体在用于训练大多数共指系统的 OntoNotes 数据集中代表性不足。Zhao 等人(2018a)提出了一种克服这种偏见的方法：他们生成了第二个性别交换的数据集，其中将 OntoNotes 中的所有男性实体都替换为女性，反之亦然，并使用合并后的原始和交换后的 OntoNotes 数据对共指系统进行重新训练，也使用去除偏见的 GloVE 嵌入(Bolukbasi 等人，2016)。所得的共指系统不再对 WinoBias 数据集表现出偏见，而不会显著影响 OntoNotes 共指精度。在后续论文中，Zhao 等人(2019)表明，在 ELMo 上下文化词向量表示和使用它们的共指系统中也存在相同的偏见。他们表明，通过数据增强对 ELMo 进行再训练可以再次减少或消除 WinoBias 共指系统中的偏见。

Webster 等人(2018)引入了另一个数据集 GAP，以及性别代词消解的任务，该任务是为性别代词开发改进的共指算法的工具。GAP 是一个带有性别歧义代词的 4,454 个句子的性别平衡标签语料库(相比之下，英语 OntoNotes 训练数据中只有 20％的性别代名词是女性的)。这些案例是通过借鉴 Wikipedia 页面上自然出现的句子而创建的，以创建难以解决的案例，该案例具有两个相同性别的命名实体和一个可能指代这两个人中的一个(或两个都不是)的歧义代词，如下所示：

(21.67) In May, Fujisawa joined Mari Motohashi' s rink as the team's skip,
moving back from Karuizawa to Kitami where **she** had spent her junior days.

Webste 等人(2018)表明，在 GAP 中，现代共指算法在解析女性代词上的表现显著逊于男性代词。Kurita 等人(2019)表明，基于 BERT 语境化单词的表示显示了类似的偏见。

21.10. 总结

本章介绍了**共指消解**的任务。

•这是一项将文本中**所指对象**和**提及**连接在一起的任务，即在**话语模型**中指代同一**话语实体**，从而形成一组**共指链**(也称为**聚类**或**实体**)。

•提及可以是**有定** NP 或**不定** NP，代词(包括**零代词**)或名称。

•实体提及的表面形式与其**信息状态**(新、旧或可推断)以及实体的**可访问性**或**显著性**如何相关联。

•一些 NP 不是**指示语**，例如在 it is raining 中使用**赘述词** it。

•许多语料库有人类标签的共指注释，可以用于监督学习，包括：**OntoNotes** 为英语、汉语、Arabic 语；ARRAU 为英语；**AnCora** 为西班牙语和加泰罗尼亚语。

•提及检测可从所有名词和命名实体开始，再使用**回指性**分类器或**指称性**分类器过滤掉未提及的内容。

•三种常见的共指架构是**提及配对**、**提及排名**和**基于实体**，每一种都可以利用基于特征或神经分类器。

•现代的共指系统倾向于端到端，在单一的端到端架构中执行提及检测和共指。

•算法学习文本跨度和中心词的表示，并学习将回指词跨度与候选先行词跨度进行比较。

•共指系统的评估是通过使用精度/召回率指标如 MUC、B3、CEAF、BLANC 或 LEA 与金牌实体标签进行比较。

•**Winograd 模式**挑战问题是困难的共指问题，此问题似乎需要世界知识或复杂的推理来解决。

•共指系统显示**性别偏见**，该偏见可以使用数据集诸如 Winobias 和 GAP 进行评估。

21.11. 文献和历史说明

自 1970 年代以来，共指一直是自然语言理解的一部分(Woods 等人 1972； Winograd 1972)。话语模型和共指的实体中心基础由 Karttunen(1969)(在第三届 COLING 会议上)提出，在语言语义学中也起着作用(Heim 1982，Kamp 1981)。但是，Bonnie Webber(1978)的论文和后续工作(Webber,1983)探索了该模型的计算方面，从而提供了对实体在话语模型中如何表示以及它们如何许可后续指代的方式的基本见解。她提供的许多例子继续挑战今天的参考理论。

Hobbs 算法是一种树搜索算法[10]，它是基于句法的长 Hobbs 算法系列中的第一个树形搜索算法，用于在自然出现的文本中稳健地标识指代。Hobbs 算法的输入是要解析的代词，连同直到和包含当前句子的的句法(成分)解析。该算法的细节取决于所使用的语法，但是由于 Kehler 等人(2004 年)的观点，可以从简化版本中了解。只是搜索当前句子和先前句子中的 NP 列表。这种简化的 Hobbs 算法按以下顺序搜索 NP：（i）在当前句子中从右到左，从代词左边的第一个 NP 开始；(ii)在前一句中从左至右；(iii)从左至右先于两个句子；(iv)在当前句子中从左到右，从代词右边的第一个名词组开始(用于加泰罗尼亚语)。在数量、性别和人称方面与代词相符的第一个名词组被选为先行词(Kehler 等人，2004)。

Lappin 和 Leass(1994)是一个有影响力的基于实体的系统，它使用权重来成分语法和其他功能，此后不久，Kennedy 和 Boguraev(1996)对其进行了扩展，该系统无需进行完整的语法分析。大约同一时期，Brennan 等人(1987)将中心化(Grosz 等人，1995)应用于代词回指的消解。随后进行了各种各样的工作，聚焦于中心化用于共指(Kameyama 1986，Di Eugenio 1990，Walker 等人 1994，Di Eugenio 1996，Strube 和 Hahn1996，Kehler1997a，Tetreault 2001，Iida 等人 2003)。Kehler 和 Rohde(2013)展示了如何将中心化与连贯性驱动的代词解释理论整合在一起。参见第 22 章，了解如何使用中心化来衡量话语连贯。

共指竞赛的一部分 US DARPA-sponsored MUC 会议提供早期标签共指数据集(1995 MUC-6 和 1998 MUC-7 语料库),为大部分以后的工作定下了基调,将社区吸引到有监督的机器学习和度量诸如 MUC 度量(Vilain 等人 1995)。后来的 ACE 评估产生了英语、中文和阿拉伯语标签的共指语料库，这些语料被广泛用于模型训练和评估。

这项 DARPA 的工作影响了社区从 90 年代中期开始的监督学习(Connolly 等人 1994；Aone 和 Bennett 1995； McCarthy 和 Lehnert 1995)。Soon 等人(2001)提出了一系列基本特征，并由 Ng 和 Cardie(2002b)

[10] 这是 Hobbs(1978)提出的两个算法中最简单的一个。

进行了扩展，并在接下来的 15 年中遵循了一系列机器学习模型。这些通常分别聚焦于代词回指的消解(Kehler 等人 2004，Bergsma 和 Lin 2006)，完整的 NP 共指(Cardie 和 Wagstaff 1999，Ng 和 Cardie 2002b，Ng 2005a)和有定 NP 的指代(Poesio 和 Vieira 1998，Vieira 和 Poesio 2000)以及分别的回指检测(Bean 和 Riloff 1999，Bean 和 Riloff 2004，Ng 和 Cardie 2002a，Ng 2004)或单例检测(de Marneffe 等人 2015)。

从成对提及到成对提及排名方法的转变是由 Yang 等人(2003)和 Iida 等人(2003)首创的，他们提出了成对提及排名方法，然后由 Denis 和 Baldridge(2008)扩展，他们提出通过 softmax 对所有先前提及的内容进行排名。当时有一个主意：在一个端到端模型中联合进行提及检测、回指和共指。这个主意源自 Ng(2005b)的早期建议，就是使用虚拟的先行词进行提及排名，允许“无所指”成为如下三个内容的一个选择：共指分类器、Denis 和 Baldridge(2007)的联合系统(把回指分类器概率与共指概率相结合)、Denis 和 Baldridge(2008)的排名模型。Rahman 和 Ng(2009)建议在单一目标下联合训练这两个模型。

简单的基于规则的共指系统在 2010 年代重新流行，部分原因是它们能够以高精度的方式对基于实体的特征进行编码(Zhou 等人 2004； Haghighi 和 Klein 2009； Raghunathan 等人 2010； Lee 等人 2011，Lee 等人 2013，Hajishirzi 等人 2013)，但是最终他们无能力去处理语义需求，即正确地处理常见名词共指的情况。

回到有监督的学习导致提及排名模型的许多进步，这些模型也扩展到了神经体系结构中，例如，使用强化学习直接优化共指评估模型(Clark 和 Manning 2016a)，进行了端到端共指跨度提取的方式(Lee 等人 2017，Zhang 等人 2018)。神经模型也被设计用来利用全局实体级别信息(Clark 和 Manning 2016b，Wiseman 等人 2016，Lee 等人 2018)。

共指也与第 23 章中讨论的实体连接任务有关。通过提供更多可能的表面形式来帮助连接到正确的 Wikipedia 页，共指可帮助实体连接，反之，实体连接可帮助提高共指消解。Hajishirzi 等人(2013)举例：

(21.68) [Michael Eisner]$_1$ and [Donald Tsang]$_2$ announced the grand opening of [[Hong Kong]$_3$ Disneyland]$_4$ yesterday. [Eisner]$_1$ thanked [the President]$_2$ and welcomed [fans]$_5$ to [the park]$_4$.

将实体连接整合到共指中可以帮助汲取百科全书知识(例如 Donald Tsang 是总统的事实)，以帮助消除对总统的提及的歧义。Ponzetto 和 Strube(2006)(2007)以及 Ratinov 和 Roth(2012)表明，从 Wikipedia 页面提取的此类属性可用于建立更丰富的共指实体提及模型。最近的研究显示了如何共同进行连接和共指联合(Hajishirzi 等人 2013，Zheng 等人 2013)，甚至还可以与命名实体标记一起进行连接(Durrett 和 Klein，2014)。

正如我们介绍的那样，共指任务涉及一个简化的假设，即回指和其先行词之间的关系是同一身份：两个共指提及是指代相同的话语指称对象。在真实文本中，这种关系可能会更加复杂，其中话语对象的不同方面可能会被抵消或重新聚焦。例子(21.69)(Recasens 等人，2011)显示了一个**隐喻(metonymy)**的示例，其中首都 Washington 通过隐喻被用来指代 US(美国)。(21.70-21.71)显示了其他示例(Recasens 等人，2011)：

(21.69) a strict interpretation of a policy requires **The U.S.** to notify foreign dictators of certain coup plots ... **Washington** rejected the bid ...

(21.70) I once crossed that border into Ashgh-Abad on Nowruz, the Persian New Year. In the South, everyone was celebrating **New Year**; to the North, **it** was a regular day.

(21.71) In France, **the president** is elected for a term of seven years, while in the United States **he** is elected for a term of four years.

有关这些共指复杂性的进一步语言学讨论，请参见 Pustejovsky(1991)，van Deemter 和 Kibble(2000)，Poesio 等人(2006)，Fauconnier 和 Turner(2008)，Versley(2008)和 Barker(2010)。

Ng(2017)提供了一个有用的关于机器学习模型在共指消解中的紧凑历史。有三种优秀的关于回指/回指消解的像书籍那么长的概述，涵盖了不同的时期：Hirst(1981)(早期工作到 1981 年左右)，Mitkov(2002)(1986-2001)和 Poesio 等人(2016)(2001-2015)。

本教材 2000 年第一版的话语章节是 Andy Kehler 写的，我们把它作为第二版章节的起点，在第三版的共指章节中还保留了 Andy 的一些可爱的散文。

21.12. 练习

(无)