

文章编号: 1000-5641(2019)05-0016-20

共指消解技术综述

陈远哲, 匡俊, 刘婷婷, 高明, 周傲英

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 共指消解旨在识别指向同一实体的不同表述, 在文本摘要、机器翻译、自动问答和知识图谱等领域有着广泛的应用. 然而, 作为自然语言处理中的一个经典问题, 它是一个 NP-Hard 的问题. 本文首先对共指消解的基本概念进行介绍, 对易混淆概念进行解析, 并讨论了共指消解的研究意义及难点. 本文进一步归纳梳理了共指消解的发展历程, 将共指消解从技术层面划分为若干阶段, 并介绍了各个阶段的代表性模型, 探讨了各类模型的优缺点, 其中着重介绍了基于规则、基于机器学习、基于全局最优化、基于知识库和基于深度学习的模型. 接着对共指消解的评测会议进行介绍, 对共指消解的语料库和常用评测指标进行解释和对比分析. 最后, 指出了当前共指消解模型尚未解决的问题, 探讨了共指消解的发展趋势.

关键词: 共指消解; 自然语言处理; 全局优化; 知识库; 深度学习

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3969/j.issn.1000-5641.2019.05.002

A survey on coreference resolution

CHEN Yuan-zhe, KUANG Jun, LIU Ting-ting, GAO Ming, ZHOU Ao-ying

(School of Data Science and Engineering, East China Normal University, Shanghai 200062, China)

Abstract: Coreference resolution is the task of finding all expressions that point to the same entity in a text; this technique is widely used for text summarization, machine translation, question answering systems, and knowledge graphs. As a classic problem in natural language processing, it is considered NP-Hard. This paper first introduces the basic concepts of coreference resolution, analyzes some confusing concepts related thereto, and discusses the research significance and difficulties of the technique. Then, we summarize research advances in coreference resolution, divide them into stages from a technical standpoint, introduce the representative approaches for each stage, and discuss the advantages and disadvantages of various methods. The summarized approaches are five-fold: rule-based, machine learning, global optimization, knowledge base, and deep learning. Next, we introduce benchmark conferences for the problem of coreference

收稿日期: 2019-07-29

基金项目: 国家重点研发计划(2016YFB1000905); 国家自然科学基金(U1811264, 61877018, 61502236, 61672234); 上海市科技兴农推广项目(T20170303)

第一作者: 陈远哲, 男, 硕士研究生, 研究方向为自然语言处理与知识图谱.

E-mail: yzchen@stu.ecnu.edu.com.

通信作者: 高明, 男, 教授, 博士生导师, 研究方向为教育计算、知识图谱、知识工程、用户画像、社会网络挖掘、不确定数据管理. E-mail: mgao@dase.ecnu.edu.cn.

resolution; in this context, we explain and compare their corpus and common evaluation metrics. Finally, this paper highlights the open problems for coreference resolution, and discusses trends and directions of future research.

Keywords: coreference resolution; natural language processing; global optimization; knowledge base; deep learning

0 引言

随着互联网的迅速发展,大量的 Web 数据不断产生,形成了一个庞大的语料库.然而,大多数 Web 上的文本仅是人类可以理解的,却难以被计算机所理解,其主要原因是大多数 Web 文本对计算机来说是缺乏语义的.近年来,通过运用自然语言处理 (Natural Language Processing) 相关技术,使得计算机也能够理解海量 Web 数据中的语义信息.自然语言处理已成为一个研究热点,如命名实体识别、实体消歧、共指消解和关系抽取等技术^[1].其中,作为自然文本理解基础的共指消解技术^[2]被广泛应用于文本摘要 (Text Summarization)、机器翻译 (Machine Translation)、自动问答 (Question Answering)、知识图谱 (Knowledge Graph) 等领域^[3].共指消解作为自然语言处理中最难的问题之一,其效果极大地影响了机器对于自然语言的理解能力.一个好的共指消解模型,能够让计算机从语料中获取更多的信息,从而大大改善自然语言后续处理的效果.

在自然文本中,经常出现同一个实体的不同表述.例如,“[陈奕迅],英文名 [Eason Chan], 1974 年 7 月出生于香港. [他]是当今华语乐坛的当红歌手.”这句话中, [陈奕迅]、[Eason Chan]、[他]这 3 个表述都指向现实生活中“香港歌手陈奕迅”这一实体.共指消解正是为识别一段文本中指向同一个实体 (Entity) 的不同表述 (Mention) 而提出的一项技术^[4].这里提到的实体是一个比较抽象的概念,在广义上讲,它对应着一个现实世界中的本体 (Ontology),在狭义上讲,它等同于知识库中的一个概念节点 (Concept Node)^[5].而表述是指文本中指代某个实体的词或短语,如名称、代词、缩写等.

如果将共指消解的过程脱离实体库而进行,那么“判断两个表述是否指向同一实体”这个过程就可以简化为“判断一个表述是否指向另一个表述”.其中,定义指出的表述为照应语 (Anaphor),定义被指向的表述为先行语 (Antecedent)^[6].根据照应语和先行语的形态以及位置关系,可以将共指分为 4 种类型:回指 (Anaphora)、预指 (Cataphora)、名词短语共指 (Coreferring Noun Phrases)、先行语分指 (Split Antecedents)^[5].表1对它们进行了辨析.

表 1 四种共指类型示例

Tab. 1 Examples of four coreference types

共指类型	定义	例子	解释
回指	照应语为人称代词,出现在先行语后面的共指情况	[小强]在平时乐于助人,因此[他]在班级中的口碑很好.	[他]是人称代词,出现在名词短语[小强]后面
预指	照应语为人称代词,出现在先行语前面的共指情况	“[我]这次彻底的失败了.”[刘总]无奈地摇头说道.	[我]是人称代词,出现在名词短语[刘总]前面
名词短语共指	照应语和先行语都是名词短语,而非人称代词的情况	2010年公布的数据显示,[中国]在第二季度已经超越日本,成为了[世界第二大经济体].	[中国]和[世界第二大经济体]都是名词短语
先行语分指	一个照应语同时对应多个先行语的组合的情况	[梅西]和[C罗]都是世界顶级的球员,[他们]惺惺相惜.	先行语[梅西]与[C罗]之和与照应语[他们]共指

共指消解的难点主要在于: 1) 共指消解已被证明是一个 NP-Hard 问题^[7], 无法在多项式时间内求得最优解; 2) 自然语言的场景和句式千变万化, 各种不同的话语可能表述了相同的语义, 而相同的话语在不同的语境下表达的含义也可能不同, 因此很难构建完整的语言学系统将共指消解的所有情况都考虑周全; 3) Web 语料的质量较低, 大多数语料均为非结构化文本, 且数据不一致和缺失的情况时有发生, 这大大影响了共指消解模型的性能。

本文的第 1 节首先对共指消解问题进行形式化的阐述, 并对一些相关的易混淆概念进行辨析; 第 2 节总结了共指消解技术的演进历史, 主要介绍了基于规则、机器学习、全局优化、知识库和深度学习这 5 个阶段的代表性方法, 并分析了这些方法的优缺点; 第 3 节介绍了共指消解在不同时期的代表性会议和语料库, 并总结了共指消解的常用评价指标及其优缺点; 第 4 节指出了共指消解目前还存在的一些问题, 并且讨论了共指消解今后的研究方向; 最后, 本文在第 5 节对全文进行了总结。

1 共指消解基本概念

1.1 形式化表示

表述的共指关系是一种等价关系, 因此共指消解的过程就是等价类划分的过程. 相互共指的表述属于同一个等价类, 不共指的表述则属于不同等价类. 由于等价关系是满足自反性、对称性、传递性的二元关系, 因此对于表述集合 M 中的表述 m_1, m_2, \dots, m_N , 显然有:

1. 自反性: m_i 与自身共指;
2. 对称性: 若 m_i 与 m_j 共指, 那么 m_j 与 m_i 也共指;
3. 传递性: 若 m_i 与 m_j 共指, m_j 与 m_k 共指, 那么 m_i 与 m_k 共指.

给定语料文本 T , 假设 T 中有 N 个表述, 第 i 个表述为 m_i , T 中所有表述构成的集合为 $M = \{m_1, m_2, \dots, m_N\}$, 而共指消解旨在寻找集合 M 的最优划分. 集合 M 的任意一个划分方案对应着唯一的共指局面, 同一个划分中的表述指向同一个实体, 不同划分中的表述指向不同实体. 假设一个划分方案将集合 M 划分成 k 个子集合 S_1, S_2, \dots, S_k , 那么这些子集合应当是全无遗漏又相互排斥的, 即满足以下约束:

1. $S_i \neq \emptyset (1 \leq i \leq k)$;
2. $S_i \cap S_j = \emptyset (1 \leq i, j \leq k, i \neq j)$;
3. $S_1 \cup S_2 \cup S_3 \cup \dots \cup S_k = M$.

可以证明, 对于包含 N 个元素的集合, 其划分的方案数随着 N 的增长呈指数增长, 而寻找最优的集合划分是 NP-Hard 问题^[7]. 因此目前所有的共指消解方法都是退而求其次, 对问题进行简化建模, 从而能够在合适的时间范围内求出一个较好的近似解.

1.2 相关概念辨析

共指消解 (Coreference Resolution) 作为自然语言处理中的一个重要课题, 近三十多年来受到许多学者的关注. 在一些早期文献中, 共指消解还有许多相近的表述, 如实体解析 (Entity Resolution)、实体匹配 (Entity Matching)、实体对齐 (Entity Alignment) 等. 其中实体解析与共指消解的定义基本相同, 而实体匹配和实体对齐则主要侧重于判断不同数据源之间的表述是否共指.

刘峤等人指出, 共指消解与命名实体识别和实体消歧同属于实体链接中的一部分^[1]. 它们

的大致过程和关系如图 1 所示.

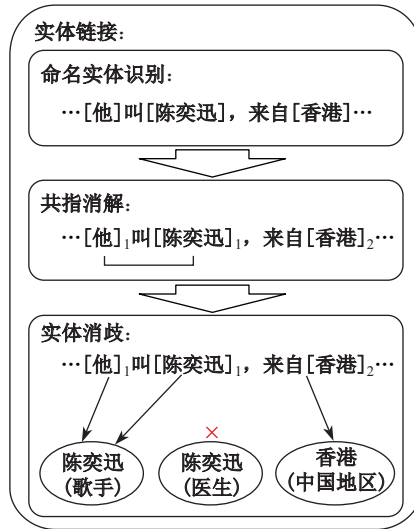


图 1 共指消解相关概念辨析

Fig. 1 Discrimination of concepts related to coreference resolution

命名实体识别 (Named Entity Recognition) 的任务是将文本中的表述识别出来, 有时命名实体识别也被叫作表述侦测 (Mention Detection). 命名实体识别技术目前较为成熟, 主流的方法大多基于序列标注方法, 代表性方法有 LSTM-CRF 模型^[8].

针对实体库中存在多个与表述同名的实体, 实体消歧 (Entity Disambiguation) 旨在消除歧义, 使得不同表述能够指向正确的实体. 例如对于[苹果]这个表述, 可能对应“苹果公司”这个实体, 也可能对应“苹果(水果)”这个实体^[9].

实体链接 (Entity Linking) 是将文本中的表述正确地链接到实体库中对应实体的过程^[10]. 实体链接的流程分为三个步骤^[1]: 1) 利用命名实体识别, 识别出文本中的表述; 2) 利用共指消解和实体消歧, 对文本中的表述进行共指划分, 并指向正确的实体; 3) 确认每个表述对应的知识库中正确实体后, 将它们链接起来.

此外, 早期的一些相关文献中还出现了回指消解 (Anaphora Resolution) 的概念^[11-12]. 回指消解与共指消解的主要区别在于, 回指消解考虑的是同一篇章中的照应语与上文中的先行语之间的语义关联性, 其不一定是等价关系; 而共指消解仅考虑两个表述指向同一个实体的情况, 其中只存在等价关系^[7].

2 共指消解研究现状

本节主要讨论共指消解的发展历程以及代表性的方法. 目前看来, 共指消解的研究大致可以分为 5 个阶段, 每个阶段的方法之间并不是独立的, 而是针对已有方法的一系列改进. 各阶段模型的特点大致如表 2 所示.

第一阶段: 始于 1978 年, 开始出现了以句法分析为基础的基于浅层语言学规则的共指消解, 代表性方法有 Hobbs 算法^[13-14]、中心理论^[15]等.

第二阶段: 始于 1995 年, 开始出现了基于二元分类和排序学习的机器学习方法, 代表性方法有决策树模型^[16]、最大熵^[17]、支持向量机^[18]等. 此外, 该时期也有一些基于无监督学习和半监督学习的共指消解方法出现, 例如聚类^[19-20]、图划分^[21]、协同训练^[22]等.

第三阶段: 始于本世纪初, 共指消解不再局限于传统的基于局部特征的机器学习框架, 开始引入了全局最优化的模型, 代表性方法有基于整数规划^[23]、启发式筛法^[24]等.

第四阶段: 始于 2011 年前后, 共指消解开始引入开放知识库的数据作为额外的特征. 代表性方法有基于众包系统^[25]、百科知识^[26-29]等.

第五阶段: 始于近几年, 深度学习开始被用于共指消解中, 并取得了当前最好的效果. 代表性方法有基于 RNN^[30]、强化学习^[31]、End-to-end^[32]等.

表 2 共指消解各研究阶段及特点

Tab. 2 Research stages and characteristics of coreference resolution

研究阶段	开始时期	代表性方法	特点
规则方法	1978年	Hobbs算法及其改进 ^[13-14,33-34] 、中心理论 ^[15,35-36]	理解和实现比较简单; 复杂的语言学规则导致泛化能力较差.
机器学习方法	1995 年	监督方法(决策树 ^[16] 、朴素贝叶斯 ^[37] 、最大熵 ^[17] 、SVM ^[18] 、CRF ^[38])、无监督方法(聚类 ^[19-20] 、图划分 ^[21] 、EM ^[39] 、LDA ^[40])、半监督方法(协同训练 ^[22] 、多视角学习 ^[41])	通过大量数据训练模型, 使得模型的泛化性能显著提升; 模型的效果高度依赖于特征工程; 模型没有考虑全局的依赖和矛盾, 效果存在一定局限性.
全局最优化方法	本世纪初	整数规划 ^[23] 、矛盾消解 ^[42] 、模式发现 ^[43] 、多通道筛法 ^[24,44-45] 、隐结构 ^[46-51] 、singleton 探测 ^[12,52-53]	基于全局最优策略, 使得模型的全局效果得到很大提升.
基于知识库的方法	2011 年	众包系统 ^[25] 、百科知识 ^[26-29]	引入开放知识作为额外特征, 很大程度避免了“知识匮乏”导致的预测错误.
深度学习方法	2016 年	前馈神经网络 ^[54-55] 、神经语言模型 ^[56] 、强化学习 ^[31] 、End-to-end ^[32] 、ELMo ^[57] 、Coarse-to-fine ^[58]	采用深度学习技术, 大大增加了模型的深层语义学习能力和泛化性能.

2.1 基于规则的方法

2.1.1 Hobbs 算法

Hobbs 算法^[13]于 1978 年由 Hobbs 提出, 是最早的共指消解算法之一. Hobbs 算法有两个不同版本: 一种是完全基于句法知识的, 被称作朴素 Hobbs (Naïve Hobbs) 算法; 还有一种是朴素 Hobbs 算法的改进版, 在原算法的基础上额外加入了语义知识^[8].

朴素 Hobbs 算法是基于纯规则的算法, 其大致流程如下: 先对文本进行句法分析, 构建出文本的句法分析树. 之后先固定一个照应语, 然后在句法分析树上从照应语节点开始按照一系列规则进行反复地回溯和广度优先遍历, 直至找到先行语.

后来有许多学者对 Hobbs 算法进行改进, 例如 Haghghi 和 Klein 在 Hobbs 算法中加入了丰富的句法语义知识, 提升了算法效果^[33]. Converse 首次将 Hobbs 算法运用在中文的共指消解中, 并针对中文语法加入了额外的约束信息^[34].

2.1.2 中心理论

中心理论 (Center Theory) 于 1995 年由 Grosz 等人提出^[15], 其中的理论核心“焦点转移”于 1981 由 Sidner 提出^[35]. 该理论最初用于预测句子焦点, 后来开始逐渐地用于代词的共指消解, 并不断受到学者的关注.

中心理论采用了不同的方法对共指消解问题进行建模, 大致思路是跟踪文本中实体的焦点变化. 中心理论认为表述应当具有局部连贯性, 因此通过在上述两种中心结构中构造规则来保持表述的连贯性, 最后就能得到合理的共指消解方案. 按照这个思路, Brennan 等人提出了 BFP 算法^[36], 算法中构造了一系列基于中心表的规则, 用于实现中心理论的局部连贯性要求.

中心理论作为一种理论模型, 自提出后被许多学者实例化. 在一系列基于中心理论的算法

被提出后, 中心理论的有效性才得以证实^[59-60]. 然而由于中心理论也是基于固定的规则, 因此同 Hobbs 算法一样缺乏泛化能力. 此外中心理论只能判断两个相邻表述之间共指与否, 从而致使模型预测能力差.

2.2 基于机器学习的方法

2.2.1 监督学习

对于共指消解问题的监督学习方法, 一般分为以下 4 种基础模型框架.

1) 表述对模型 (Mention-pair Model) 将共指消解问题看作表述对的二元分类问题, 是最常见的一类模型. 该类模型中分类器根据表述对的上下文特征以及距离特征, 判定表述对共指与否^[61-62]. 然而这种模型表示存在两种缺陷: 1. 只关注先行语和照应语之间的关系, 却忽略了先行词两两之间的相互关系; 2. 表述对的特征有时候不足以判断是否共指, 可能存在代词语义过空、表述性别难以分辨等种种情况.

2) 表述排序模型 (Mention-ranking Model) 将共指消解问题看作排序学习问题^[63]. 该类模型需要训练一个打分器, 对于一个照应语, 能够将前 k 个先行语与照应语之间按照共指可能性进行打分, 并按照分值进行排序. 该模型是双候选词共指模型 (Twin-candidate Coreference Model)^[64]的一种扩展, 或者说双候选词共指模型是 $k = 2$ 的特例. 由于该模型同时考虑了多个先行语之间的排序关系, 因此弥补了 1) 中提到的表述对模型的第一个缺陷.

3) 实体表述模型 (Entity-mention Model)^[65] 将共指消解问题看作实体与表述的二元分类问题, 这里的实体就是共指的先行语集合. 由于一个实体包含多个共指先行语, 它们的上下文特征信息能够互补, 因此弥补了 1) 中提到的表述对模型的第二个缺陷.

4) 实体排序模型 (Entity-ranking Model/Cluster-ranking Model) 结合了模型 2)、3)^[18]. 给定一个照应语, 该模型首先需要先行语集合按照共指关系进行划分, 将先行语集合转化为实体集合. 然后该模型再将这些实体按照与照应语的共指可能性进行打分排序. 该模型既考虑了多个实体之间的排序关系, 又实现了实体先行语的特征互补, 同时克服了 1) 中表述对模型的两个缺陷.

基于这 4 种基础模型框架, 许多基于机器学习的共指消解方法开始出现.

McCarthy 和 Lehnert 基于表述对模型, 提出了二元分类的决策树 (Decision Tree) 模型^[16]. 该模型采用了 C4.5 决策树, 通过计算各个特征选项的信息增益比, 来确定决策树的结构. 该模型具有很好的可解释性, 但是当数据集发生轻微扰动时, 决策树容易产生大幅度变化.

Ge 等人提出了基于朴素贝叶斯 (Naïve Bayes) 的共指消解模型^[37]. 该方法假设各个表述对之间的共指与否是相互独立的, 原理简单, 但是由于其条件独立性假设与实际不符, 导致模型的精度不足.

Ponzetto 和 Strube 使用最大熵模型 (Maximum Entropy) 进行共指消解^[17]. 该模型使用最大熵模型进行表述对二分类, 将已有的表述对特征和共指局面作为特征函数(即已发生的事实), 构造出使得条件熵最大的关于表述对是否共指的 0-1 条件分布. 该模型在理论上较为完美, 但是在实际中训练计算量很大.

Rahman 等人使用了支持向量机 (Support Vector Machines) 进行共指消解, 并试验了其在 4 种模型框架中的性能^[18]. 最终评测结果表明, 基于实体排序模型框架的系统性能是最优的. Rahman 等人在随后的论文中指出^[66], 在数据集和特征基本相同的情形下, 4 种 baseline 的性能从高到低依次为: 实体排序模型 > 表述排序模型 > 实体表述模型 > 表述对模型.

McCallum 和 Wellner 将共指消解看作序列标注问题, 采用条件随机场 (Conditional Random Field) 来预测标签序列^[38]. 该模型考虑文本中表述构成的序列, 对表述序列进行实体编号标注, 这样实体编号相同的表述就是共指的. 条件随机场与隐马尔科夫模型 (Hidden Markov Models) 相比, 没有了观测独立性假设, 因此能够在不考虑表述之间依赖的情况下合并原输入中

的大量特征,使得标注更加准确.

2.2.2 无监督学习

在实践中,相比较于海量的无标注文本,由于人工标注的成本高昂,往往带标注的数据集规模都较小.因此许多学者提出了无监督的共指消解方法,将大规模的容易获取的无标注数据用于模型训练,也取得了不错的效果.

Cardie 和 Wagstaff 利用无监督的表述聚类算法进行共指消解,将聚到同一个簇中的表述看作是共指的^[19].周俊生等人提出了基于图划分的无监督共指消解方法^[21].该方法将文本中的表述看做无向图中的节点,表述之间的相似度看做节点之间的无向边.该模型引入了模块度 (Modularity) 的概念,用于衡量划分的合理性^[67].谢永康等人提出了基于谱聚类的共指消解方法^[20],谱聚类体现了“类内距离最小,类间距离最大”的原则,有效地提高了表述划分的准确率.

Ng 采用基于 EM 算法 (Expectation Maximization Algorithm) 的聚类进行共指消解^[39].该模型将表述的共指方案看做隐变量,不断循环执行 E 步和 M 步直至模型参数收敛. E 步计算给定当前参数下似然函数的期望; M 步计算使期望对数似然函数最大化的参数.通过多轮迭代发现使得似然函数最大的共指方案.

Bhattacharya 和 Getoor 使用了 LDA (Latent Dirichlet Allocation) 模型进行共指消解,并且在此基础上提出了改进的 LDA-ER 模型^[40].LDA-ER 在传统 LDA 模型的基础上,增加了噪声模型 (Noise Model),用来进一步提升模型的效果. LDA-ER 模型能够求出每个表述对应的实体分布,从而根据实体分布来判断每个表述分别指向哪个实体.该模型在语料库规模很大的条件下,能够取得较好的效果.

2.2.3 半监督学习

半监督学习是介于监督学习和无监督学习之间的一类方法,其优势在于既能利用带标注数据保证模型的精度,又能通过无标注数据提升模型的泛化能力.由于基于半监督学习的共指消解模型能够利用较小的成本获得较高的性能,因此越来越受到学界的重视.

Muller 等人采用了基于协同训练 (Co-Training) 的半监督共指消解模型,大大减少了人工标注量^[22,68]; Raghavan 等人在医学概念的共指消解中分别采用了基于协同训练和多视角学习 (Multi-view Learning) 的半监督方法^[41].在多视角学习中,该模型具体使用了后验正则化 (Posterior Regularization) 框架下的最大熵模型进行训练^[69].

2.3 基于全局最优化的方法

传统基于机器学习的共指消解虽然相较于基于规则的方法,在模型性能上得到了很大的提升,但其仍存在不足之处: 1) 训练数据的特征往往是局部的,没有考虑全局的依赖关系和语义特征; 2) 可能违背共指等价关系的传递性,比如“A 与 B 共指, B 与 C 共指,而 A 与 C 不共指”.为了有效克服这些问题,基于全局最优化的共指消解模型开始被提出.

Denis 和 Baldridge 使用整数线性规划模型对共指消解问题进行建模^[23].该模型定义了 0-1 变量 $x_{\langle i,j \rangle}$ 表示文本中第 i 个表述和第 j 个表述构成的表述对是否共指,定义了 0-1 变量 y_j 来表示第 j 个表述是否是照应语.

McCallum 和 Wellner 针对已有模型给出的共指划分不满足传递性的缺陷,使用改进的条件随机场进行表述序列标注,从而对图进行分割,将表述分割至不存在矛盾的划分中^[42].文中定义了不一致三角形 (Inconsistent Triangle),用来表示一个共指矛盾情况.不一致三角形越多,共指矛盾越多.另外文中采用不一致检查函数 (Inconsistency-checking Function) 来度量三个表述对应的实体是否构成不一致三角形,并将其也作为条件随机场的特征函数,赋予其权值.该模型可以权衡不一致三角形的出现个数,有效地降低了共指消解违反等价性的可能.

Yang 和 Su 则提出从语料库中自动寻找有效的模式 (Patterns) 进行共指消解的方法^[43].所

谓的模式, 就是两个共指表述以及它们之间经常出现的短语共同构成的“模板结构”, 例如“[表述 A]的全名叫[表述 B]”、“[表述 A]被人们亲切称呼为[表述 B]”等. 与传统的基于人工规则不同, 本模型能自动从语料库中统计各个模式的频率 (Frequency) 特征和可靠性 (Reliability) 特征, 对语料库中出现的各个模式进行打分和排序. 预测表述对是否共指时, 只需将表述对之间的文本与模式库进行比较, 即可预测出表述对的共指概率.

Raghunathan 等人提出的基于筛子 (Sieve) 的多通道筛法 (Multi-pass Sieve Approach)^[24]. 该方法首先给定一个包含所有表述对的集合 U , 然后通过一系列的筛子将不共指的表述对从集合 U 中剔除. 后来 Chen 等人将这种筛法推广到了中文和阿拉伯文^[44], 不同的语言对应的筛子略微不同. 基于该方法构建的共指消解系统在 CoNLL-2011 的评测中获得了英语共指消解的最好成绩^[45].

Fernandes 等人提出了隐共指树 (Latent Coreference Trees) 模型用于共指消解^[46]. 对于文本中的每个表述, 该模型能够预测出下文中与其共指可能性高的其他表述, 并将这些表述作为其子节点. 对于每个表述进行预测后, 就能生成一个森林, 森林中的每棵树就是一个共指簇. 在基础特征之上, 该模型引入了 C4.5 决策树的思想, 在隐共指树的分叉中加入了基于熵变的派生特征^[47]. 由于共指树可以视作隐结构, 因此该模型采用了隐结构 SVM (Latent Structural SVM)^[48]进行隐共指树的预测.

Daume 等人在隐共指树结构的基础上延迟进行 LaSO (Learning as Search Optimization)^[49]直至文档末尾, 训练效果要优于基于 Early update 的 LaSO 策略^[50]. 之后 Martschat 和 Strube 总结出了隐结构训练和预测框架, 并且将表述对模型、表述排序模型、共指树模型分别定义了其隐结构, 代入框架进行训练^[51]. 实验结果表明三种模型加入了隐结构后, 表述排序模型的提升最为明显.

Recasens 等人发现语料库中一句话通常包含多个实体, 但是每个实体很少被多次提及, 也就是一句话中的表述之间很少出现共指. 基于这个先验知识, Recasens 构建了一个生存期 (Lifespan)^[52]分类器, 用于区分话语中的一个表述是单独表述 (Singleton) 还是共指表述 (Coreferent)^[53]. 由于单独表述一定不和其他表述共指, 对被预测为单独表述的表述则无需参与后续的共指消解过程. Moosavi 和 Strube 在该模型的基础上, 利用更加简单有效的特征, 对单独表述检测的搜索空间进行了剪枝, 使得模型取得了更好的效果^[70].

Wiseman 等人在表述排序模型的基础上, 将共指消解问题拆分成两个子任务: 照应语检测和先行语排序^[71]. 照应语检测也可以看做是单独表述的检测, 在 Recasens 的生存期模型^[52]和 Ma 的贪心算法^[72]中已经被提出过, 该模型主要通过训练一个打分函数来预测表述是照应语的概率. 对于先行语排序问题, 模型也训练了一个先行语打分函数用于共指概率排序. 此外模型在这两个子任务中分别引入了相应的预训练模型, 提升了模型的性能.

2.4 基于知识库的方法

人们的一些先验知识有助于共指消解任务, 然而由于先验知识获取比较困难, 导致有利于共指消解的先验知识是比较匮乏的. 例如, [菠萝]和[凤梨]具有共指关系, 而传统的模型却无法有效利用这种先验知识. 为了克服这个困难, 基于知识库的共指消解模型开始出现. 这类模型从知识库中提取额外的特征, 能够发现表述之间的一些隐含关系, 从而提高共指消解的性能^[26].

Vesdapunt 等人提出了基于众包的共指消解算法, 有效利用众包知识来对传统共指消解模型进行补充^[25]. 该模型先采用机器学习方法训练共指消解模型, 生成以表述为节点, 表述共指概率为边的一个概率无向图. 然后该模型构造了一个面向用户的标注系统, 服务器根据策略生成表述对, 用户只需判断给定表述对是否共指, 并提交表单即可. 该系统根据共指的传递性、以及基于概率图的启发式方法来指定表述对生成策略, 最大程度地减少系统与用户的交互次数, 利用

尽可能少的用户标注, 将共指的概率图进行补充完善. 因此该模型能够在提高共指消解准确率的前提下, 尽可能地降低知识的获取成本.

Rahman 和 Ng 充分利用了百科数据, 从中抽取知识来提升基础共指消解模型的性能^[26]. 该模型主要抽取了 3 种数据源的知识: 大规模知识库、带共指标注数据、无标注数据. 其中大规模知识库包括了 YAGO^[73]和 FrameNet^[74], YAGO 中可以抽取出类似“(Einstein, MEANS, Albert Einstein)”的三元组, FrameNet 中可以根据指定谓词对来提取共指表述对; 带共指的标注数据是经过人工标注的语句, 从中可以获得到名词对特征和动词对特征; 无标注数据是通过启发式方法和解析树, 从语料库中自动识别出的(姓名, 指代)二元组. 该模型以表述对模型和实体排序模型为基础, 加入知识库特征后, 准确率和召回率均得到大幅提升.

Ratinov 和 Roth 在多重筛法模型的基础上, 将 Wikipedia 页面中的消歧项和额外的关键词作为额外特征, 有效地改善了多重筛法模型的性能^[27]. Durrett 和 Klein 等人在语言特征提取上下功夫, 并在此基础上从知识库中结合了浅语义特征来改进共指消解模型的效果^[28]. Sorulaze 等人对共指消解模型采用误差分析技术, 通过引入知识库来弥补误差对共指消解性能造成的影响^[29]. 该模型主要采用了 Wikipedia 的实体链接来丰富表述的特征, 并且利用了 Wikipedia 页面中的别名项和 WordNet 中的同义词项作为额外的筛子, 有效地提高了共指消解模型的性能.

2.5 基于深度学习的方法

近年来, 诸如 Word Embedding^[75]、LSTM^[76]、Attention^[77]等深度学习组件在自然语言处理各领域展现出了巨大潜力, 深度学习的方法也渐渐被用于共指消解问题的求解. 得益于深度学习模型的强大泛化能力, 基于深度学习的共指消解模型也取得了巨大的性能改进, 其关键是深度学习能够通过多层级的网络层对数据进行多层级的抽象表示, 也就是所谓模型的“深度”^[78].

Park 等人在表述对模型的基础上, 结合了多通道筛法^[24]和深度学习, 用于韩语的共指消解^[54]. 该模型首先对文本表述进行多通道筛法, 将其输出作为神经网络的输入, 网络结构为全连接层构成的前馈神经网络, 网络输出单元为预测表述对是否共指的 softmax 单元. 该模型取得了韩语共指消解的最好效果.

Wiseman 等人在表述排序模型的基础上, 采用循环神经网络 (RNN) 来更好地捕获共指消解的全局特征^[30]. 该模型在 RNN 中使用了 LSTM 单元 (Long Short-Term Memory unit)^[76], 使得表述的全局特征能够通过记忆细胞 (Memory Cell) 序列长距离传递下去.

Clark 和 Manning 构建了 4 个深度学习模块用于共指消解^[55]. 表述对编码器 (Mention-pair Encoder) 负责将输入的表述对特征向量进行编码; 实体对编码器 (Cluster-pair Encoder) 负责池化再编码; 表述排序模型负责预训练和搜索空间的剪枝; 实体排序模型负责生成最后的共指结果^[49,79]. 神经网络的训练策略采用 RMS-Prop^[80], 并且加入了 L2 正则化和 Dropout 防止网络的过拟合^[81].

同年, Clark 和 Manning 在之前工作的基础上, 又提出了一个重要观点: 共指消解的过程中应当要考虑不同共指决策的重要性. 为了让模型能够判断不同决策的重要性, 避免模型做出严重错误的决策, 两人在之前深度学习模型^[55]的基础上加入了强化学习 (Reinforcement Learning) 策略^[31]. 该模型只保留了表述对编码器和表述排序模型这两个模块, 强化学习部分采用了策略梯度法 (Policy Gradient Algorithm)^[82].

Ji 等人提出了一种新的神经语言模型 ENTITYNLM^[83], 用于改进之前的共指消解模型的性能. ENTITYNLM 在神经语言模型^[56]的基础上, 对每个词添加了额外随机变量来表示词与实体表述的关系, 并在时间点 t 上添加了新的变量更新规则, 用于动态的实体表示.

Lee 等人构造了端到端 (End-to-end) 的神经网络共指消解模型, 在不需要句法分析和命名实体识别的情况下超越了过去所有模型的效果^[32]. 该模型创造性地使用了 span-ranking 方法,

直接对 span-pair 进行打分排序 (span 就是文本的子串, 长度为 n 的句子有 $n(n-1)/2$ 个 span). 端到端网络结构的嵌入层使用了 GloVe Embedding^[84] 和半监督 Embedding^[85], 特征抽取层使用了双向 LSTM, 注意力机制 (Attention)^[77] 等重要技术. 模型的目标就是希望能够预测句子中最优的 span 划分, 即对应着一个好的共指消解局面.

随着深度上下文词向量 ELMo 的出现, Peters 等人首次将其加入到端到端神经网络共指消解模型中^[57], 通过动态生成词向量, 克服了传统词向量技术无法解决的“一词多义”的问题, 大大提升了共指消解的性能. 同年, Lee 等人又在端到端神经网络^[32]的基础上, 引入了由粗到细 (Coarse-to-fine) 的推断策略, 对于某个表述的所有先行语, 先用简单的打分函数求出共指概率最高的前 M 个先行语, 然后仅对这 M 个先行语采用复杂的打分函数^[58]. 该方法目前取得了共指消解任务的最好效果.

3 共指消解国际评测

随着共指消解不断受到工业界和学术界的关注, 公开、公平、标准的评测方法显得尤为重要. 目前共指消解的国际评测主要包括两部分: 语料库和评测指标. 一个高质量、统一的语料库, 才能够使得不同系统之间不会因为语料库的不同而导致模型之外的性能差异. 同样地, 一个好的评测指标, 才能够更好地衡量模型的真实性能.

3.1 共指消解会议及语料库

表 3 展示了与共指消解相关的会议及其公布的语料库, 其中某些会议只有在特定的年份才发布共指消解的任务.

表 3 共指消解会议及语料库

Tab. 3 Conferences and corpus of coreference resolution

会议名称	举办时间	共指消解任务年份	语料库	特点
MUC	1987-1997	1995、1998	MUC数据集	主题只与军事、科技相关, 只包含英文
ACE	2000-2008	2003-2008	ACE数据集	包含新闻专线、广播、报纸中语料, 首次加入中文
TAC	2008-至今	2009-2017	TAC数据集	取代了ACE会议, 共指消解任务开始过渡到基于维基百科的实体链接任务
SemEval	1998-至今	2010	OntoNotes2.0数据集	没有将单独表述(Singleton)标注出来, 增加了共指消解的难度
CoNLL	1999-至今	2011、2012	OntoNotes4.0数据集 OntoNotes5.0数据集	OntoNotes4.0(CoNLL 2011)只支持英文, OntoNotes5.0(CoNLL 2012)中加入了中文和阿拉伯文, 是目前最经典的数据集

消息理解系列会议 MUC(Message Understanding Conferences) 是最早包含有共指消解任务的会议, 会议于 1987 年由美国国防高级研究计划委员会 DARPA(Defense Advanced Research Projects Agency) 创建. 该会议的主要目的是促进信息抽取领域的发展. 从 1987 年至 1997 年, MUC 系列会议一共举办了 7 届(MUC-1 至 MUC-7), 语料库的主题主要与军事、科技相关. 从第六届会议 MUC-6^[86] 开始, 会议中加入了命名实体识别和共指消解的任务, 其语料库语言只包含英文.

自动内容抽取 ACE(Automatic Content Extraction) 评测会议开始于 2000 年, 由美国国家

标准与技术研究院 NIST(National Institute of Standards and Technology) 举办, 一直持续到了 2008 年. ACE旨在为促进人类自然语言的自动化处理, 并于 2003 年开始加入了共指消解任务, 其中首次加入了中文共指消解的语料库^[87]. 该语料库中主要包含了新闻专线、广播、报纸中的语料.

文本分析会议 TAC(Text Analysis Conference) 开始于 2008 年, 一直举办至今. TAC 也是由 NIST 所举办, 取代了之前的 ACE. 共指消解任务出现于 2009、2010、2011、2017 年的 TAC 会议, 其形式变成了 KBP(Knowledge Base Population) 任务.

语义评估 SemEval(Semantic Evaluation) 评测会议开始于 1998 年, 一直举办至今. SemEval 早期每 3 年举办一届, 2010 年之后每隔 1 至 2 年举办一届. SemEval 于 2010 年加入了多语言共指消解任务, 语料库为 OntoNotes2.0 数据集^[88]. 与之前的数据集不同, OntoNotes 数据集中没有将孤立表述标注出来, 只标注了发生共指关系的表述, 这也一定程度上增加了共指消解的难度(共指消解过程开始需要命名实体识别的预处理).

计算自然语言学习 CoNLL(Computational Natural Language Learning) 评测会议开始于 1999 年, 每年举办一次. CoNLL 于 2011 年举办了英文共指消解评测, 采用了 OntoNotes4.0 数据集^[89]. 该会议每年举办一次, 由 SIGNLL(Special Interest group on Natural Language Learning) 负责组织. CoNLL 于 2012 年再次举办了共指消解评测, 并且将语料库更新至了 OntoNotes5.0 数据集^[90]. 该数据集提供了英文、中文、阿拉伯文 3 种语言用于评测多语言共指消解, 目前成为共指消解任务中最经典的数据集.

3.2 共指消解评测指标

一个共指消解局面对应着表述集合的一个划分, 因此对一次共指消解的预测结果, 我们通过比较其预测划分和实际划分的差异可以衡量预测结果的好坏. 一般地, 我们定义测试集中表述的真实划分为 Key, 定义模型预测出的输出划分为 Response, 那么一个评价指标就可以形式化地表示成一个函数 $\text{Score} = f(\text{Key}, \text{Response})$. 表 4 是语料库中某个句子共指情况的真实划分和两个系统对其的预测划分^[76].

表 4 共指划分的一个例子

Tab. 4 An example of coreference partition

[鲍勃] ₁ 今天计划出去玩, 于是[他] ₂ 打电话叫[查理] ₃ 一同前往[海滩] ₄ . 然而, [查理] ₅ 没有回应[他] ₆ 的呼叫, 因为[他] ₇ 已经在[海滩] ₈ 了.	
Key:	{1,2,6}鲍勃, {3,5,7}查理, {4,8}海滩
Response 1:	{1,2,6,7}鲍勃, {3,5}查理, {4,8}海滩
Response 2:	{1,2,3,5,6,7}鲍勃/查理, {4, 8}海滩

我们还可以采用共指链图来表示一个共指局面. 其中共指的表述按照编号(在文本中出现位置)从小到大进行排列, 用箭头进行连接. 对于表 4, 可以画出对应的共指链图, 如图 2 所示.

由于共指消解可以看做是对等价类的划分, 因此同一条共指链上的表述两两之间都是共指的, 无需显式地画出所有共指对. 此外, 共指链图和共指集合划分可以进行等价转换. 评价共指消解结果的好坏, 可以从集合划分差异的角度进行, 也可以从共指链图差异的角度进行.

共指消解评测中最常用的指标有 MUC-score、ACE-value、B-CUBED、CEAF, 以及比较新的 BLANC、LEA. 由于同一个模型在不同的评测指标下的性能可能会有所不同, 因此大多数的共指消解模型会在准确率 (Precision), 召回率 (Recall), F1-score 等指标下进行评测, 这样才

能够更加全面地评估模型的综合性能. 表 5 展示了各评测指标的特点.

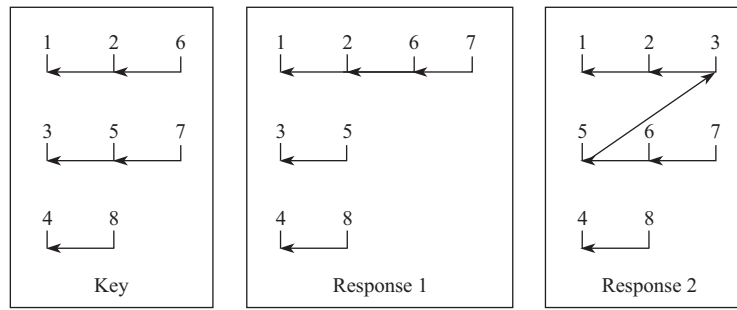


图 2 共指链图的一个例子

Fig. 2 An example of coreference linking graph

表 5 共指消解评测指标

Tab. 5 Evaluation metrics of coreference resolution

指标名称	关注点	优/缺点
MUC-score ^[91]	主要统计Key和Response中共指的共指链接个数	计算方法较为简单; 但是无法计算所有表述均为单独表述的情况, 且对错误严重程度不同的共指链同等看待
ACE-value ^[87]	除了考察共指链预测, 还考察了实体和表述类型的预测正确与否	将实体类型也考虑到评测指标中; 但是与当前标准共指消解任务有一定区别, 因此只适用于ACE数据集, 现在很少采用该指标
B-CUBED ^[92]	从划分的角度, 直接对表述进行逐个统计	克服了 MUC-score 的缺点; 但是当 Key 中所有表述共指时召回率一定为100%, 当Key中都是单独表述时准确率一定为 100%, 这显然是错误的
CEAF ^[93]	建立了 Key 到 Response 之间共指链的一对一映射, 可看做二分图匹配问题	克服了 B-CUBED 的缺点; 但是其忽视了 Response 中正确但未匹配的共指链, 并且忽视了共指集合的大小
BLANC ^[94-95]	同时考虑了共指表述对和非共指表述对的准确率和召回率, 最后求其平均	克服了CEAF的缺点, 是一种较新的评测指标; 但是由于该指标对单独表述是否识别过于敏感, 没有被广泛采用
LEA ^[96]	以 Key 和 Response 的共指交集的链接数为基础, 再按照共指集合大小加权	克服了 BLANC 的缺点, 是一种较新的评测指标, 同时考虑了共指链的完整性和共指集合的大小; 但是由于提出较晚, 暂未被广泛使用, 且相较先前方法运算略为复杂

3.2.1 MUC-score 指标

MUC-score 最初用于评判 MUC 数据集, 在 MUC 会议停办之后, 该指标仍然在其他会议的数据集上继续沿用下去^[91]. 设 Key 和 Response 中共同出现的共指链接有 C 个, 那么 MUC-score 定义该 Response 的准确率为 C 除以 Response 中共指链接个数, 召回率为 C 除以 Key 中共指链接的个数. 一条共指链接在共指链图中可以是单个箭头或者多个箭头的串联, 例如图 2 中 Key 里的 $\langle 1, 2 \rangle$, $\langle 1, 6 \rangle$ 都可以看做是一条共指链接. 但是在一个共指局面对应的共指链图中, 每个表述只能出现在一条共指链接的末端, 因此如果将 $\langle 2, 6 \rangle$ 看做一条共指链接, 那么 $\langle 1, 6 \rangle$ 就将被排除.

MUC-score 计算简单, 作为一个重要的性能衡量指标, 广泛应用于各个共指消解系统的评测环节. 但是 MUC-score 存在一些缺点: 1) 当句子中所有的表述均为单独表述时(即不存在共指链接), 准确率和召回率的分子均为零, MUC-score 无法计算; 2) MUC-score 对错误严重程度不同的共指链同等看待, 因此不容易检测出模型的致命错误.

3.2.2 ACE-value 指标

ACE-value 用于 ACE 评测会议的性能评价,除了考虑表述共指链的预测结果之外,该指标还需要考虑表述识别是否正确,以及预测的表述和实体的类型是否正确^[87]. Response 的 ACE-value 的值等于 1 减去其错误率,而错误率主要受两个因素影响:实体类型(例如人名、地名、机构名)预测错误、表述类型(例如名字、名词、代词)预测错误.当 Response 和 Key 完全一致时,ACE-value 能达到 100%;当 ACE-value 为 0% 时,不一定是最坏情况,可能是句子中根本没有共指表述.此外,ACE-value 可以小于 0.

由于 ACE-value 是针对 ACE 会议的评测任务而提出的,其中涉及了实体类型预测的部分,与现在标准的共指消解任务有一定的区别,因此 ACE-value 已经很少被采用.

3.2.3 B-CUBED 指标

B-CUBED 是一种常用的共指消解评测指标,它克服了 MUC-score 存在的缺点^[92].与 MUC-score 统计共指链接不同,B-CUBED 是从划分的角度直接对表述进行统计,因此可以适用于不存在共指链接的情况.B-CUBED 依次对每个表述都计算准确率和召回率,再进行加权求和得到总体的准确率和召回率.其中每个表述的权值一般为表述数量的倒数,也可以根据错误程度的不同为每个表述分配不同的权值.

一个 Response 的 B-CUBED 具体计算方法如下:对于第 i 个表述,设在 Key 中与其共指的表述集合为 UK_i ,在 Response 中与其共指的表述集合为 UR_i ,那么有

$$Precision_i = \frac{|UK_i \cup UR_i|}{|UR_i|},$$

$$Recall_i = \frac{|UK_i \cup UR_i|}{|UK_i|}.$$

对于该 Response,有

$$Precision = \sum_{i=1}^N w_i \times Precision_i,$$

$$Recall = \sum_{i=1}^N w_i \times Recall_i.$$

B-CUBED 也存在一些问题,例如 Key 中所有表述均在同一条共指链上时召回率一定为 100%,Key 中没有共指链接时准确率一定为 100%.不过即便如此,目前,B-CUBED 仍然是一种主流的共指消解评测指标,具有很高的参考价值.

3.2.4 CEAF 指标

CEAF(Constrained Entity-Alignment F-Measure) 指标的计算较为复杂,但是克服了 MUC-score 指标和 B-CUBED 指标存在的缺点^[93].CEAF 与过去的指标之间最主要的区别在于,其建立了 Key 到 Response 之间共指链的一对一的映射.

首先定义 Key 和 Response 中共指链集合分别为 R 和 S ,其中共指链分别为 R_i 和 S_j .然后定义 ϕ 函数来度量两条共指链的相似性,为了简单起见,我们直接设其为两条共指链中共同表述的个数,即:

$$\phi(R_i, S_j) = |R_i \cap S_j|.$$

设 R 中有 n 条共指链, S 中有 m 条共指链, 那么 R 的共指链到 S 的共指链的映射方法可以看做是二元组的集合 $g \in G$. 例如 $g = \{\langle 1, 2 \rangle, \langle 2, 1 \rangle\}$ 表示 R_1 映射到 S_2 , R_2 映射到 S_1 . CEAF 需要求出相似度最高的共指链映射方法 g^* , 即:

$$g^* = \operatorname{argmax}_{g \in G} \sum_{\langle i, j \rangle \in g} \phi(R_i, S_j).$$

该问题可以看做是求解 $n \times m$ 条边的带权二分图的最大匹配问题, R 和 S 中的共指链分别为二分图两端的节点, R 和 S 的共指链之间的 ϕ 函数即为节点之间的边权值. 带权二分图的最大匹配求解有成熟的算法, 一般实践中采用 Kuhn-Munkres 算法^[97-98]. 求得最大匹配后, 该 Response 的准确率和召回率即可计算如下:

$$Precision = \frac{\sum_{\langle i, j \rangle \in g^*} \phi(R_i, S_j)}{\sum_i \phi(S_i, S_i)},$$

$$Recall = \frac{\sum_{\langle i, j \rangle \in g^*} \phi(R_i, S_j)}{\sum_i \phi(R_i, R_i)}.$$

现在我们用 CEAF 对图 2 例子求其 CEAF, 可画出对应的带权二分图, 如图 3 所示.

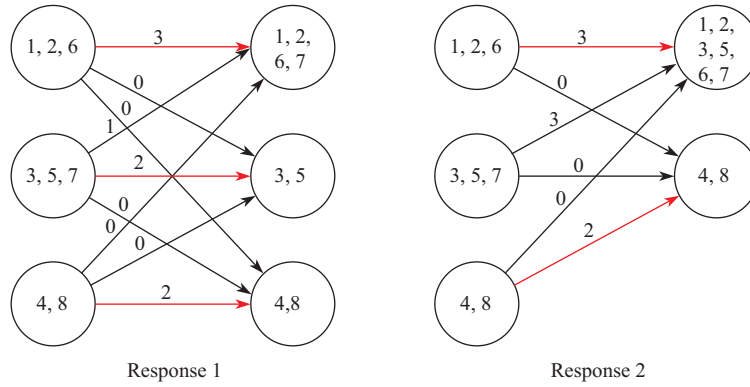


图 3 带权二分图最大匹配的例子

Fig. 3 Example of the maximum matching of the weighted bipartite graph

可以算出, Response1 的最大匹配值为 7, Response2 的最大匹配值为 5. 进一步可以算出 Response1 的准确率为 $7/8$, 召回率为 $7/8$; Response 的准确率为 $5/8$, 召回率为 $5/8$. 需要注意, 由于本文采用了较简单的 ϕ 函数, 导致了 Response 的准确率和召回率总是相等, 如果采用更复杂的 ϕ 函数, 可以避免该问题. 由于 CEAF 指标较为合理, 因此现在也成为了主流的共指消解评测手段之一. 不过 CEAF 也存在缺陷, 其忽视了 Response 中未被匹配但正确的共指链, 并且忽视了共指集合的大小.

3.2.5 BLANC 指标

BLANC(BiLateral Assessment of Noun-phrase Coreference) 指标同时统计了共指表述对和非共指表述对^[94-95]. 该指标分别定义 C_k 和 C_r 为 Key 和 Response 中的共指表述对集合, N_k 和 N_r 为 Key 和 Response 中的非共指表述对集合, 那么有:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|},$$

$$R_n = \frac{|N_k \cap N_r|}{|N_k|}, P_n = \frac{|N_k \cap N_r|}{|N_r|}.$$

最后, BLANC 指标的召回率和准确率就是对上式求平均值, 即召回率 = $(R_c + R_n)/2$, 准确率 = $(P_c + P_n)/2$.

BLANC 属于一种比较新的评测指标, 但是并没有被广泛使用, 原因是该指标存在一个缺陷: 对于系统是否识别出单独表述很敏感. 因为若一个系统识别出了单独表述, 那么会增加很多非共指表述对(因为该单独表述和其他表述都不共指), 导致评测结果与识别不出单独表述的系统产生分歧.

3.2.6 LEA 指标

LEA(Link-based Entity-Aware Metric) 是目前最新的评测指标, 它克服了传统评测指标中的缺陷^[96]. 假设 Key 中的每个共指集合为 k_i , Response 中的每个共指集合为 r_j , $\text{link}(x)$ 表示共指集合 x 中的链接数(含有 n 个表述的集合有 $n(n-1)/2$ 条链接). 那么 LEA 的召回率和准确率的计算如下:

$$\text{Recall} = \frac{\sum_{k_i \in K} \left(|k_i| \times \sum_{r_j \in R} \frac{\text{link}(k_i \cap r_j)}{\text{link}(k_i)} \right)}{\sum_{k_z \in K} |k_z|},$$

$$\text{Precision} = \frac{\sum_{r_i \in R} \left(|r_i| \times \sum_{k_j \in K} \frac{\text{link}(r_i \cap k_j)}{\text{link}(r_i)} \right)}{\sum_{r_z \in R} |r_z|}.$$

LEA 由于近两年才被提出, 因此使用该指标进行评测的文献还比较少. LEA 的一个很大优势在于, 其既考虑了共指链的完整性、又考虑了共指集合的大小. 因此 LEA 既能根据共指集合的规模进行重要性加权, 也不会对表述的识别过于敏感.

4 研究趋势与展望

4.1 尚未解决的难题

共指消解研究至今, 新的方法层出不穷, 并且很多方法都取得了很好的效果. 但是目前共指消解的性能仍然不够理想, 即便当前性能最好的模型^[32]的平均 F1 值也仅为 68.8%. 共指消解之所以还有这么大的提升空间, 很重要的原因在于以下问题还尚未被解决.

1) 模型缺乏语义推理的能力

Peng 等人提出, 共指消解存在许多判断困难的情况, 这些情况需要系统能够深入理解文本的语义和上下文才能够做出正确的判断^[99]. 例如, “[斑马]₁ 在草原中遇见了 [狮子]₂……[它]₂ 开始撕咬[它]₁ 的后背.”这句话中, 当前的共指消解模型很难判断出句中的两个[它]分别指向哪个实体. 而人类由于有先验知识“狮子和斑马是捕食者与被捕食者的关系, 而撕咬一般是捕食者对被捕食者的行为”, 因此能够推断出是“狮子撕咬斑马”. 对人来说很容易的推理, 目前的共指消解模型却很难做到.

2) 缺乏共指消解的语料库

相较于其他 NLP 问题, 共指消解的公开语料库资源极其有限. 一方面原因在于, 共指消解任务在早期没有形成统一规范, 导致在一些诸如回指消解的相似任务上设计的语料库无法适配于当今的共指消解任务; 另一方面原因在于, 共指消解的任务定义本身也经历过了多次的更改, 不同版本语料库之间的格式差异巨大, 导致共指消解系统很难同时兼容多种语料库. 在当今的共指消解任务中, 使用最广泛的语料库为 OntoNotes 系列数据集^[89-90].

3) 模型效果过于依赖前置模型的性能

事实上, 对前置模型的高度依赖是 NLP 相关领域中长期存在的通病. 在共指消解模型中, 分词、词性标注、句法分析、命名实体识别、词嵌入等模型的效果好坏, 会高度影响共指消解的效果. 即便是两个架构完全相同的共指消解系统, 若上述某个环节存在性能差异, 也会导致最终共指消解的性能不同. 虽然当前一些 End-to-end 的模型已经在一定程度上缓解出了问题 (Lee 的系统^[32]无需对句子进行命名实体识别和句法分析), 但问题仍未被完全解决.

4.2 未来的发展趋势

1) 采用知识图谱抽取开放特征

当前已有学者尝试引入开放知识库的知识来辅助共指消解, 并且取得了不错的效果^[26-29]. 但是目前对知识的利用, 还仅仅是简单地将百科页面固定位置的信息作为额外特征. 近年来随着开放知识图谱的出现, 通过知识图谱来改善共指消解的语义理解能力将成为可能. 通过知识图谱进行知识推理, 可以对表述的额外特征进行精确的、动态的查询; 通过在知识图谱上进行表述之间的路径查询, 能够将路径中蕴含的知识作为表述对的额外特征.

2) 更为充分地利用无标注数据

当前的共指消解模型还未能充分利用无标注数据, 无监督方法和半监督方法在近几年的共指消解任务中也很少出现. 理论上, 在待标注数据规模有限的情况下, 半监督方法能够大大增加模型的泛化性能, 提高模型的准确率和召回率^[100]. 因此对于数据集相对匮乏的共指消解任务, 半监督方法的研究也是未来的重要趋势. 此外, 随着 Pseudo label^[101]、半监督 ladderNet^[102]等半监督深度学习技术的出现, 将共指消解、深度学习和半监督学习三者进行结合也成为了可能.

3) 强化学习展现用武之地

随着 AlphaGo^[103]的横空出世, 其内部的强化学习技术取得的巨大成功引起了学术界和工业界的极大重视. Clark 等人首次将强化学习用于共指消解问题^[31], 并且实验结果证明了强化学习能够有效提升共指消解的性能. 通过强化学习, 共指消解系统能够判断不同决策的重要程度, 从而避免做出严重错误的决策. 对于一个训练参数已经收敛的共指消解系统, 在其基础上继续进行强化学习, 能够让模型效果进一步提升. 因此, 如何从强化学习的角度更好地对共指消解问题进行建模, 是将来的重要研究方向.

4) 更完备的 End-to-end 模型

当前, 深度学习在 NLP 领域取得了很大的成功, 其中一类 End-to-end 的模型更是大放异彩. 例如 End-to-end 的文本摘要和情感分类^[104]、End-to-end 的机器翻译^[105]、以及 End-to-end 的共指消解等模型^[32], 都在各自的领域取得了突破性的成果. End-to-end 模型最大的优势在于, 其充分地发挥了神经网络强大的学习能力, 避免了传统“流水线”模块之间的级联误差. 另外, End-to-end 模型使得人们不用像传统模型那样过多地关注语言学细节, 降低了共指消解的研究门槛. 因此, 如何设计精度和集成度更高的 End-to-end 模型, 将会是共指消解领域热门的研究课题.

5 总 结

共指消解是自然语言处理中的重要研究问题, 从上世纪 70 年代至今, 该研究问题经历了长足的发展. 客观来看, 共指消解的研究进程顺应了人工智能的发展趋势, 现已进入了海量知识背景的深度学习时代. 本文对共指消解的本质进行了深入的剖析, 从方法论的角度对各阶段的共指消解模型进行了分析评判, 并着重对一些代表性模型进行了概述. 此外本文系统地梳理了共指消解的语料库和评价指标, 指出了共指消解的研究现状和发展趋势, 为未来共指消解的相关研究奠定了基础.

[参 考 文 献]

- [1] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(3): 582-600.
- [2] 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.
- [3] GETOOR L, MACHANAVAJJHALA A. Entity resolution: Theory, practice & open challenge [J]. Proceedings of the Very Large Data Bases Endowment, 2012, 5(12): 2018-2019.
- [4] MELLI G, ESTER M. Supervised identification and linking of concept mentions to a domain-specific ontology [C]//Proceedings of the 19th ACM International Conference on Information & Knowledge Management. 2010: 1717-1720.
- [5] JURAFSKY D, MARTIN H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [M]. New Delhi: Pearson Education, 2000.
- [6] LANG J, QIN B, LIU T, et al. Intra-document coreference resolution: The state of the art [J]. Journal of Chinese Language and Computing, 2008, 17(4): 227-253.
- [7] 宋洋, 王厚峰. 共指消解研究方法综述 [J]. 中文信息学报, 2015, 29(1): 1-12.
- [8] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of NAACL-HLT. 2016: 260-270.
- [9] 高艳红, 李爱萍, 段利国. 面向实体链接的多特征图模型实体消歧方法 [J]. 计算机应用研究, 2017, 34(10): 2909-2914.
- [10] LI Y, WANG C, HAN F Q, et al. Mining evidences for named entity disambiguation [C]// Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining. 2013: 1070-1078.
- [11] DEEMTER K V, KIBBLE R. On coreferring: Coreference in MUC and related annotation schemes [J]. Computational Linguistics, 2000, 26(4): 629-637.
- [12] MITKOV R. Anaphora resolution: The state of the art [D]. Wolverhampton: University of Wolverhampton, 1999.
- [13] HOBBS J R. Resolving pronoun references [J]. Journal of Lingua, 1978, 44: 311-338.
- [14] WALKER M A. Evaluating discourse processing algorithms [C]// Proceedings of the 27th Annual Meeting of Association of Computational Linguistics. Vancouver, 1989.
- [15] GROSZ B, JOSHI A, WEINSTEIN S. Centering: A framework for modelling the local coherence of discourse [J]. Journal of Computational Linguistics, 1995, 21(2): 203-225.
- [16] MCCARTHY J, LEHNERT W. Using decision trees for coreference resolution [C]// Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995.
- [17] PONZETTO S P, STRUBE M. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution [C]// Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2006: 192-199.
- [18] RAHMAN A, NG V. Supervised models for coreference resolution [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 968-977.
- [19] CARDIE C, WAGSTAFF K. Noun phrase coreference as clustering [C]// Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora. 1999: 277-308.
- [20] 谢永康, 周雅倩, 黄萱菁. 一种基于谱聚类的共指消解方法 [J]. 中文信息学报, 2007, 21(2): 77-82.
- [21] 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.
- [22] MULLER C, RAPP S, STRUBE M. Applying co-training to reference resolution [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 352-359.
- [23] DENIS P, BALDRIDGE J. Joint determination of anaphoricity and coreference resolution using integer programming [C]// Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. 2007: 236-243.
- [24] RAGHUNATHAN K, LEE H, RANGARAJAN S, et al. A multi-pass sieve for coreference resolution [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010.
- [25] VESDAPUNT N, BELLARE K, DALVI N. Crowdsourcing algorithms for entity resolution [C]// Proceedings of the VLDB Endowment. 2014: 1071-1082.
- [26] RAHMAN A, NG V. Coreference resolution with world knowledge [C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 814-824.
- [27] RATINOV L, ROTH D. Learning-based Multi-Sieve Co-Reference Resolution with Knowledge [M]. Association for Computational Linguistics, 2012: 1234-1244.
- [28] DURRETT G, KLEIN D. Easy Victories and Uphill Battles in Coreference Resolution [M]. Association for Computational Linguistics, 2013: 1971-1982.
- [29] SORALUZE A, ARREGI O, ARREGI X, et al. Enriching basque coreference resolution system using semantic knowledge sources [C]//Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes. Association for Computational Linguistics, 2017: 8-16.

- [30] WISEMAN S, RUSH A M, SHIEBER S M. Learning global features for coreference resolution [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.
- [31] CLARK K, MANNING C D. Deep reinforcement learning for mention-ranking coreference models [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2256-2262.
- [32] LEE K, HE L H, LEWIS M, et al. End-to-end neural coreference resolution [C]// Conference on Empirical Methods in Natural Language Processing. 2017: 188-197.
- [33] HAGHIGHI A, KLEIN D. Simple coreference resolution with rich syntactic and semantic features [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 1152-1161.
- [34] CONVERSE S P. Pronominal Anaphora Resolution in Chinese [D]. Pennsylvania: University of Pennsylvania, 2006.
- [35] SIDNER C. Focusing for interpretation of pronouns[J]. Computational Linguistics. 1981, 7(4): 217-231.
- [36] BRENNAN S E, FRIEDMAN M W, POLLARD C. A centering approach to pronouns [C]// Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics. 1987: 155-162.
- [37] GE N Y, HALE J, CHARNIAK E. A statistical approach to anaphora resolution [C]// Proceedings of the ACL 1998 Workshop on Very Large Corpora. 1998.
- [38] MCCALLUM A, WELLNER B. Conditional models of identity uncertainty with application to noun coreference [C]// International Conference on Neural Information Processing System. 2004: 905-912.
- [39] NG V. Unsupervised models for coreference resolution [C]// Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008: 640-649.
- [40] BHATTACHARYA I, GETOOR L. A latent Dirichlet model for unsupervised entity resolution [C]// SIAM International Conference on Data Mining. 2006.
- [41] RAGHAVAN P, FOSLERLUSSIER E, LAI A M. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features [C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 731-741.
- [42] MCCALLUM A, WELLNER B. Conditional models of identity uncertainty with application to noun coreference [C]// Proceedings of Neural Information Processing Systems. 2004: 905-912.
- [43] YANG X, SU J. Coreference resolution using semantic relatedness information from automatically discovered patterns [C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007: 528-535.
- [44] CHEN C, NG V. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution [C]// Joint Conference on EMNLP & CONLL-Shared Task. Association for Computational Linguistics, 2012: 56-63.
- [45] LEE H, PEIRSMAN Y, CHANG A, et al. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task [C]// Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task. 2011: 28-34.
- [46] FERNANDES E R, SANTOS C N, MILIDIU R L. Latent trees for coreference resolution[J]. Computational Linguistics, 2014, 40(4): 801-835.
- [47] FERNANDES E R, MILIDIU R L. Entropy-guided feature generation for structured learning of Portuguese dependency parsing [C]// Computational Processing of the Portuguese Language. 2012: 146-156.
- [48] YU C N J, JOACHIMS T. Learning structural SVMs with latent variables [C]// Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 1169-1176.
- [49] DAUME H, MARCU D. Learning as search optimization: Approximate large margin methods for structured prediction [C]// Proceedings of the 22nd International Conference on Machine Learning. 2005: 169-176.
- [50] BJORKELUND A, KUHN J. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 47-57.
- [51] MARTSCHAT S, STRUBE M. Latent structures for coreference resolution [J]. Transactions of the Association for Computational Linguistics, 2015(3): 405-418.
- [52] RECASENS M, MARNEFFE M C, POTTS C. The life and death of discourse entities: Identifying singleton mentions [C]// The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2013: 627-633.
- [53] MARNEFFE M C, RECASENS M, POTTS C, et al. Modeling the lifespan of discourse entities with application to coreference resolution [J]. Journal of Artificial Intelligence Research, 2015, 52: 445-475.
- [54] PARK C, CHOI K H, LEE C K, et al. Korean coreference resolution with guided mention pair model using deep learning [J]. ETRI Journal, 2016, 38(6): 1207-1217.
- [55] CLARK K, MANNING C D. Improving coreference resolution by learning entity-level distributed representations [EB/OL]. [2019-05-03]. <https://arxiv.org/pdf/1606.01323.pdf>.

- [56] MIKOLOV T, KARAFIAT M, BURGET L, et al. Recurrent neural network based language model [C]// Conference of the International Speech Communication Association. 2010: 1045-1048.
- [57] PETERS M E, NEUMANN M, LYYER M, et al. Deep contextualized word representations [C]//North American Chapter of the Association for Computational Linguistics. 2018: 2227-2237.
- [58] LEE K, HE L H, ZETTLEMOYER L. Higher-order coreference resolution with coarse-to-fine inference [C]//North American Chapter of the Association for Computational Linguistics. 2018: 687-692.
- [59] LAPPIN S, SHALOM H J. An algorithm for pronominal anaphora resolution [J]. Computational Linguistics, 1994, 20(4): 535-561.
- [60] POESIO M, STEVENSON R, EUGENIO B D, et al. Centering: A parametric theory and its instantiations[J]. Computational Linguistics, 2004, 30(3): 309-363.
- [61] NG V, CARDIE C. Improving machine learning approaches to coreference resolution [C]// Meeting of the Association of Computational Linguistics. 2002: 104-111.
- [62] PONZETTO S P, STRUBE M. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution [C]// Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. 2006: 192-199.
- [63] DENIS P, BALDRIDGE J. Specialized models and ranking for coreference resolution [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008: 660-669.
- [64] YANG X, ZHOU G, SU J, et al. Coreference resolution using competitive learning approach [C]// Proceedings of the Association of Computational Linguistics. 2003: 176-183.
- [65] YANG X F, SU J, LANG J, et al. An entity-mention model for coreference resolution with inductive logic programming [C]// Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2008: 843-851.
- [66] RAHMAN A, NG V. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution [J]. Journal of Artificial Intelligence Research, 2011, 40: 469-521.
- [67] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks [J]. Phys Rev E, 2004, 69(2): 026113.
- [68] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training [C]// Proceedings of the 11th Annual Conference on Learning Theory. 1998: 92-100.
- [69] GANCHEV K, GRACA J, GILLENWATER J. Posterior regularization for structured latent variable models [J]. Journal of Machine Learning Research, 2010, 11(1): 2001-2049.
- [70] MOOSAVI N S, STRUBE M. Search space pruning: A simple solution for better coreference resolvers [C]//Proceedings of NAACL-HLT 2016. Association for Computational Linguistics, 2016: 1005-1011.
- [71] WISEMAN S, RUSH A M, SHIEBER S M, et al. Learning anaphoricity and antecedent ranking features for coreference resolution [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015: 1416-1426.
- [72] MA C, DOPPA J R, ORR J W, et al. Prune-and-score: Learning for greedy coreference resolution [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014.
- [73] SUCHANEK F, KASNECI G, WEIKUM G. YAGO: A core of semantic knowledge unifying wordnet and Wikipedia [C]// Proceedings of the World Wide Web Conference. 2007: 697-706.
- [74] BAKER C F, FILLMORE C J, LOWE J B. The Berkeley FrameNet project [C]// Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics. 1998: 86-90.
- [75] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2019-05-10]. <https://arxiv.org/pdf/1301.3781.pdf>.
- [76] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9: 1735-1780.
- [77] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-06-02]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [78] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436.
- [79] CLARK K, MANNING C D. Entity-centric coreference resolution with model stacking [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015: 1405-1415.
- [80] HINTON G, TIELEMAN T. Lecture 6.5-RmsProp: Divide the gradient by a running average of its recent magnitude [J]. COURSE: Neural Networks for Machine Learning, 2012, 4: 26-30.
- [81] HINTON G, SRIVASTAVA N, KRIZHEVSKY I, et al. Improving neural networks by preventing coadaptation of feature detectors [EB/OL]. [2019-06-20]. <https://arxiv.org/pdf/1207.0580.pdf>.
- [82] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Machine Learning, 1992, 8(3/4): 229-256.
- [83] JI Y F, TAN C H, MARTSCHAT S, et al. Dynamic entity representations in neural language models [EB/OL]. [2019-06-10]. <https://arxiv.org/pdf/1708.00781.pdf>.

- [84] PENNINGTON J, SOCHER R, MANNING C D. GloVe: Global vectors for word representation [C]//Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.
- [85] TURIAN J, RATINOV L, BENGIO Y. Word representations: A simple and general method for semi-supervised learning [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010: 384-394.
- [86] GRISHMAN R, SUNDHEIM B. Message understanding conference-6: A brief history [C]// Proceedings of the 16th Conference on Computational linguistics. 1996: 466-471.
- [87] NIST, US. The ACE 2003 Evaluation Plan V [R]. US National Institute for Standards and Technology (NIST), 2003.
- [88] RECASENS M, MARQUEZ L, SAPENA E, et al. SemEval-2010 Task 1 OntoNotes English: Coreference Resolution in Multiple Languages [M]. Philadelphia: Linguistic Data Consortium, 2011.
- [89] PRADHAN S S, RAMSHAW L, MARCUS M, et al. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes[C]// Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning. 2011: 1-27
- [90] PRADHAN S, MOSCHITTI A, XUE N W, et al. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes [C]// Proceedings of the Shared Task of the 16th Conference on Computational Natural Language Learning. 2012: 1-40.
- [91] VILAIN M, BURGER J, ABERDEEN J, et al. A model-theoretic coreference scoring scheme [C]// Proceedings of the 6th Conference on Message Understanding. 1995: 45-52.
- [92] BAGGA A, BALDWIN B. Algorithms for scoring coreference chains [C]// Proceedings of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation. 1998: 563-566.
- [93] LUO X. On coreference resolution performance metrics [C]// Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005: 25-32.
- [94] RECASENS M, HOVY E. BLANC: Implementing the rand index for coreference evaluation[J]. Natural Language Engineering, 2011, 17(4): 485-510.
- [95] LUO X, PRADHAN S, RECASENS M, et al. An extension of BLANC to system mentions [C]// Meeting of the Association for Computational Linguistics. 2014: 24.
- [96] MOOSAVI N S, STRUBE M. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric [C]// Meeting of the Association for Computational Linguistics. 2016: 7-12.
- [97] KUHN H W. The Hungarian method for the assignment problem [J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.
- [98] MUNKRES J. Algorithms for the assignment and transportation problems [J]. Journal of the Society for Industrial & Applied Mathematics, 1957, 5(1): 32-38.
- [99] PENG H R, KHASHABI D, ROTH D. Solving hard coreference problems [EB/OL]. [2019-05-1]. <https://arxiv.org/pdf/1907.05524.pdf>.
- [100] ZHOU Z H. A brief introduction to weakly supervised learning [J]. National Science Review, 2017, 5(1): 44-53.
- [101] LEE D H. Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks [C]// International Conference on Machine Learning. 2013.
- [102] RASMUS A, VALPOLA H, HONKALA M, et al. Semi-supervised learning with ladder networks [J]. Computer Science, 2015: 1-9.
- [103] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529: 484-489.
- [104] MA S, SUN X, LIN J Y, et al. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification [C]// International Joint Conferencces on Artificial Intelligence. 2018.
- [105] CHO K, VAN MERRENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Conference on Empirical Methods in Natural Language Processing. 2014: 1724-1734.

(责任编辑: 林 磊)