基于领域知识的图模型词义消歧方法

鹿文鹏 [1,2] 黄河燕 [1] 吴 昊 [1]

摘 要 对领域知识挖掘利用的充分与否, 直接影响到面向特定领域的词义消歧 (Word sense disambiguation, WSD) 的性能. 本文提出一种基于领域知识的图模型词义消歧方法, 该方法充分挖掘领域知识, 为目标领域收集文本领域关联词作为文本领域知识, 为目标歧义词的各个词义获取词义领域标注作为词义领域知识; 利用文本领域关联词和句子上下文词构建消歧图, 并根据词义领域知识对消歧图进行调整; 使用改进的图评分方法对消歧图的各个词义结点的重要度进行评分, 选择正确的词义. 该方法能有效地将领域知识整合到图模型中, 在 Koeling 数据集上, 取得了同类研究的最佳消歧效果. 本文亦对多种图模型评分方法做了改进, 进行了详细的对比实验研究.

关键词 词义消歧, 领域知识, 图模型, 词义领域, 文本领域

引用格式 鹿文鹏, 黄河燕, 吴昊. 基于领域知识的图模型词义消歧方法. 自动化学报, 2014, **40**(12): 2836−2850

DOI 10.3724/SP.J.1004.2014.02836

Word Sense Disambiguation with Graph Model Based on Domain Knowledge

LU Wen-Peng[1,2] HUANG He-Yan[1] WU Hao[1]

Abstract Whether domain knowledge is fully utilized would impact the performance of word sense disambiguation (WSD) on a specific domain. A WSD method with graph model based on domain knowledge is proposed in the paper. The method makes full use of domain knowledge: first, the keywords related with target text domain are collected as text domain knowledge, and domain annotations of each sense of target ambiguous word are obtained as sense domain knowledge; second, a disambiguation graph is constructed with text domain knowledge and sentence context words; thirdly, the disambiguation graph is adjusted based on sense domain knowledge; finally, the sense nodes in the graph are scored with an improved evaluation method to judge the right sense. This WSD method effectively integrates domain knowledge with graph model. Evaluation is performed on Koeling dataset. Compared with similar methods, the WSD method yields state-of-the-art performance. Besides, multiple graph evaluation models are improved and compared in detail.

Key words Word sense disambiguation (WSD), domain information, graph model, sense domain, text domain

Citation Lu Wen-Peng, Huang He-Yan, Wu Hao. Word sense disambiguation with graph model based on domain knowledge. *Acta Automatica Sinica*, 2014, **40**(12): 2836−2850

自然语言中普遍存在一词多义现象. 词义消歧 (Word sense disambiguation, WSD) 指根据多义词所处的上下文环境确定其词义, 属于自然语言理解的底层研究, 对机器翻译、信息检索、文本分类、自动问答等均有直接影响[1−6]. 词义消歧属于 AI−完全 (AI-complete) 问题, 迄今为止, 一直是困扰计算语言学者的最复杂问题之一.

词义消歧分为有监督、无监督和基于知识库的方法. 有监督方法根据词义标注语料库, 利用机器学习技术训练分类器, 判定新实例的词义. 该方法消歧正确率高, 但其效果高度依赖标注语料库的规模和质量; 由于难以获得充足的标注语料, 导致其难以应用于大规模词义消歧任务. 无监督方法利用聚类算法对上下文相似的实例进行聚类, 不使用任何人工知识 (如词典、标注语料库等), 仅能区分词义类别, 无法对词义进行明确标注, 实质是一种词义辨析 (Word sense discrimination) 方法[1]. 基于知识库的方法根据歧义词所处的上下文, 利用各种知识资源 (如词典、知识本体、固定搭配等) 推测其词义. 该方法的正确率通常不及有监督的方法, 但其可利用丰

收稿日期 2014-01-21 录用日期 2014-05-01
Manuscript received January 21, 2014; accepted May 1, 2014
国家重点基础研究发展计划 (973 计划) (2013CB329303), 国家自然科学基金 (61132009), 山东省高等学校科技计划 (J12LN09) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329303), National Natural Science Foundation of China (61132009), and Shandong Province Higher Educational Science and Technology Program (J12LN09)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 北京理工大学计算机学院 北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081 2. 齐鲁工业大学理学院 济南 250353
1. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Application, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 2. School of Science, Qilu University of Technology, Jinan 250353

富的各类知识库资源, 具有较高的消歧覆盖率, 能满足大规模词义消歧任务的需求. 研究表明, 基于知识库的方法在特定领域上能取得优于有监督方法的效果[7]. 鉴于基于知识库的方法是目前唯一能真正用于大规模词义消歧任务的方法及其在 SemEval 评测中的优良效果, 该方法逐渐受到研究者的重视. 近年来, 基于知识库的词义消歧方法的研究可划分为两条路线.

路线 1 是建立或完善知识库, 提供更加完备的消歧知识, 先后出现了 WordNet Domain[8]、eXtended WordNet、Google N-gram Corpus、BabelNet[9] 等消歧知识库. 现有基于知识库的词义消歧方法倾向于寻找歧义词与句子上下文消歧特征词的语义关联来实现消歧, 多侧重于对词语相似、WordNet 语义相关、Wiki 链接信息等关联知识的利用, 往往忽视了领域知识. 已有相关研究表明, 歧义词所在文本的领域属性及歧义词自身各词义的领域属性, 对歧义词的词义具有指示作用[8]; 在特定领域文本的词义消歧上, 表现更为明显[10]. 现有的词义消歧研究工作多是面向通用领域文本 (General domain text); 在特定领域文本 (Specific domain text) 中, 歧义词的上下文环境、词义分布等均会发生显著变化, 对研究者提出了新的挑战. 特定领域的词义标注语料的获取更加困难, 导致有监督方法的性能明显下降, 难以应用; 而基于知识库的词义消歧方法亦受到领域知识不足的制约[11]. 随着面向特定领域的词义消歧研究的兴起[7, 10−11], 如何充分利用领域知识改善基于知识库的词义消歧方法的性能, 成为一个迫切需要解决的问题.

路线 2 是创建消歧模型, 深入挖掘已有知识库的内在结构化关联信息, 先后出现了词语相似度度量 (Similarity measure)、结构化语义互联 (Structural semantic interconnections)、Personalized PageRank 和 BabelNet 图消歧模型等. 其中, 图模型可有效表达词义结点之间的语义关联关系, 可将消歧问题转化为词义结点的重要度评价问题, 具有良好的消歧性能. 已有研究中, 文献 [7, 12−13] 采用单一的 PageRank 类算法对图模型中的结点进行评价; 文献 [9] 尽管采用了多种不同方法, 但均比较简单, 未能取得最优效果. 词义消歧图模型的结点评价方法尚有待进一步深入研究.

针对以上问题, 本文提出了一种基于领域知识的图模型词义消歧方法, 挖掘领域知识来改善消歧性能, 改进完善了多种图模型评分方法, 并作了详细对比研究. 本文将领域知识划分为两个层次: 一是文本领域, 即歧义词所处的句子、段落或文档的领域属性; 二是词义领域, 即歧义词各个词义的领域属性. 本文从这两个角度着手提出了一种挖掘领域知识改进词义消歧效果的有效方法. 在 Koeling 数据集[14] 上, 与已有同类研究相比[7, 14], 该方法取得了最佳的消歧效果. 本文的主要贡献体现在以下方面:

1) 提出一种基于领域知识的消歧图的构建和调整方法. 该方法通过对数似然统计发现与目标领域具有密切关系的领域关联词作为文本领域知识, 利用领域关联词与歧义词的句子上下文共同构建消歧图; 通过 WordNet Domain[8] 查找各个词义的领域标注作为词义领域知识, 对于领域标注与歧义词所在的文档领域密切相关的词义, 调整其消歧图的关联边, 以增加其重要度. 该方法可充分挖掘领域知识的作用, 改进词义消歧的效果.

2) 对多种消歧图模型评分方法做了改进和对比研究. 针对已有图评分方法的不足, 分别从关联边权重、双向路径两个角度出发, 进行改进, 并作了详细对比研究, 有效改善了词义消歧效果.

3) 首次提出以文本领域关联词作为文本领域知识, 与歧义词的句子上下文, 共同构建消歧图; 以词义领域标注作为词义领域知识, 调整消歧图; 同时考虑关联边权重和双向路径对消歧图进行评分以获得正确词义的词义消歧方法. 实验结果表明, 本文方法具有较高的消歧召回率, 可适用于特定领域的词义消歧任务.

本文第 1 节介绍词义消歧的相关工作; 第 2 节详细说明基于领域知识的图模型词义消歧方法; 第 3 节给出实验和分析; 最后对全文进行总结.

1 相关工作

近年来, 国外的 Navigli[1], 国内的卢志茂等[3]、王瑞琴等[5] 从不同角度对词义消歧进行了详细综述.

Gale 等提出了 “One sense per discourse” 的假设[15], 即在一段论述中, 当一个多义词多次出现时, 倾向于使用同一个词义. Magnini 等对 Gale 等的假设做了进一步延伸, 提出了 “One domain per discourse” 假设[8], 即在一段论述中, 当一个多义词多次出现时, 词义倾向于属于同一个领域. Magnini 等[8] 通过对 SemCor 语料库的实证统计, 验证了该假设的正确性. 根据该假设, McCarthy 等进行特定领域的词义消歧研究工作, 取得了很好的效果[16]. 这些已有研究工作说明, 领域知识对于歧义词的词义判定有重要影响; 利用领域知识来改善词义消歧

系统的性能是可行的.

领域知识对词义消歧效果的改善有促进作用, 在面向特定领域的词义消歧工作中表现更为突出. 本文以面向特定领域的词义消歧为研究对象, 探讨如何挖掘并应用领域知识改善消歧效果, 相关工作的介绍亦由此展开. 面向特定领域的词义消歧方法需要解决的关键技术问题是领域适应 (Domain adaptation) 问题, 主要涉及有监督 (Supervised adaptation) 和基于知识库的适应 (Knowledge-based adaptation) 方法[11].

有监督适应方法主要围绕寻找更为有效的消歧特征和降低词义标注工作量而展开. Agirre 等[17] 使用奇异值分解 (Singular value decomposition, SVD) 对未标注语料和源领域训练语料获得的消歧特征进行降维处理, 寻找与源领域、目标领域均密切相关的消歧特征 (SVD features); 将 SVD 特征与常用的局部搭配、句法依存、词袋等特征 (Original features) 进行整合; 实验表明, 只需取 40 % 的目标领域的训练语料累加至源领域训练语料上, 使用整合特征训练 SVM 分类器, 便可取得完全使用目标领域的训练语料所获得的分类器的消歧效果; 该方法可有效降低目标领域的词义标注工作量[17]. Chan 等提出了一种基于主动学习 (Active learning) 策略的领域适应方法, 利用主动学习机制挑选目标领域的实例进行人工标注, 减轻词义标注工作量; 并利用期望最大化 (Expectation maximization, EM) 算法来估计歧义词在目标领域上的优势词义, 结合计数合并 (Count merging) 技术进一步改进效果[18]. Zhong 等提出了一种融合特征增加和主动学习机制的领域适应方法, 在主动学习的迭代过程中使用特征增加策略, 可取得减小标注工作量并提高消歧正确率的双重效果[19]. 已有的有监督的领域适应技术的研究主要是针对扩充消歧特征和减小标注工作量而展开. 这只能减轻对训练语料的部分依赖; 面对有监督方法对训练语料的海量需求, 显然无法从根本上解决困扰有监督方法的数据稀疏 (Data sparseness) 问题.

基于知识库的领域适应方法的研究主要围绕领域消歧知识的发现和利用而展开. McCarthy 等提出了一种自动获取特定领域上优势词义 (Predominant sense) 的方法[16], 首先依据句法关系在特定领域语料中为各个歧义词查找句法分布相似度 (Distributional similarity score) 最高的一组相关词; 然后由 WordNet Similarity 工具包计算歧义词的各个候选词义与这一组相关词的相似度, 选择相似度最大的作为正确词义. Aitor 等设计了一种面向特定领域的一体化消歧系统[20], 首先利用 Tybots (Term yielding robots) 工具为各个歧义词在目标领域上收集一组领域相关词; 而后利用基于图的 Personalized PageRank 来推断正确的词义. 这两种方法均挖掘歧义词的领域相关词表作为消歧知识, 并据此为歧义词的不同消歧实例统一选择词义, 而未考虑不同消歧实例的真实上下文环境. Reddy 等提出了一种将特定领域信息融合到局部上下文环境的方法[21], 首先利用 McCarthy 的方法[16] 为各个候选词义计算在当前特定领域上的领域词义评分; 而后, 根据对数似然统计为目标领域获取领域关键字, 并为其计算领域关键字评分; 按领域词义评分和关键字评分来初始化针对局部上下文构建的消歧图中的结点和边的权重, 利用 Personalized PageRank 选择正确的词义. Stevenson 等提出了一种针对医学文献的基于领域信息的词义消歧方法[10], 为不同的医学主题, 各自收集一组主题关联词, 并进行评分; 对于出现在某特定医学主题的消歧实例, 将该医学主题的主题关联词和歧义句中的词语共同构成消歧上下文, 由 Personalized PageRank 来判断正确的词义. Reddy 等[21] 和 Stevenson 等[10] 的方法, 均同时考虑了领域信息和局部上下文环境, 利用领域信息来对局部上下文构建的消歧图进行优化调整; 整体来看, 均是对 Personalized PageRank 算法的改进应用. Personalized PageRank 是目前非常流行的一种消歧算法, 具有较好的消歧效果. 但其仅依据 WordNet 的语义关系来构建消歧图, 这种单一的知识来源不利于其知识覆盖面的扩展, 限制了其效果的提升.

Navigli 等将 WordNet 的词典知识与 Wikipedia 的百科知识进行整合, 构建了大规模多语知识网络 BabelNet[9], 避免了对 WordNet 的单一依赖. WordNet 和 Wikipedia 的相互补充, 使 BabelNet 具有较高的词汇覆盖率和丰富的语义关联. 根据 BableNet 可为歧义句构建消歧知识图, 为歧义词选择与上下文概念关联性最高的词义作为正确词义. Magnini 等采用半人工的方式, 为 WordNet 的所有词义概念均标注了领域信息, 构建了词义领域知识库 WordNet Domain, 并将其应用于词义消歧研究[8]. BableNet 与 WordNet Domain 为面向特定领域的基于知识库的词义消歧研究提供了极大便利. 本文的主要工作将基于此开展.

2 基于领域知识的图模型词义消歧方法

2.1 基于图模型的词义消歧

图模型是目前词义消歧领域的一个研究热点. 基于图模型的词义消歧方法起源于词汇链 (Lexical chain) 研究[1]. 词汇链指彼此之间存在词汇语义关联 (如 is-a, has-part 等) 的一组词. 如: eat → dish → vegetable → aubergine. 在图模型中, 词汇链可看作图中概念结点之间存在的路径关联关系.

随着词汇链研究的不断深入, Galley 等提出了图模型词义消歧方法[22], 包含两个阶段:

1) 图构建阶段. 根据词汇知识库, 获得歧义词及其上下文词对应的全部词义概念及词义概念之间存在的语义关联关系; 将词义概念作为顶点, 语义关联关系作为边, 构建消歧图.

2) 图评分阶段. 对图中各顶点的重要度进行评分, 选择评分最高、最重要的词义概念作为正确词义.

Mihalcea 等提出的基于 PageRank 的词义消歧方法[13]、Agirre 等提出的 Personalized PageRank[12]、Navigli 提出的结构化语义互连 (Structural semantic interconnections, SSI) 及 BabelNet[9] 等均属于图模型词义消歧研究.

Navigli 等构建了大规模多语知识网络 BabelNet[9], 将 WordNet 词典知识和 Wikipedia 百科知识相整合, 两类不同的知识源相互补充, 有效提高了 BabelNet 的词汇覆盖率, 丰富了概念之间的语义关联关系. BabelNet 已在多个不同数据集上进行了词义消歧测试, 取得了较好的消歧效果[9]; 且其已提供开源工具包供研究者使用. 鉴于此, 本文消歧图的创建以 BabelNet 为基础.

2.2 领域知识的界定

陈文亮等[23] 提出了一种基于领域词典的文本特征表示方法, 利用领域词典中的领域关联词和领域特征属性作为文本特征, 改进文本分类的效果; 领域关联词指带有强烈主题倾向的词语, 可用来表征一段文本的领域知识; 领域特征属性指词语所属的领域属性特征. 陈分别从两个维度, 即文本的领域关联词和词语的领域特征属性, 来描述文本特征.

受陈文亮等的工作[23] 启发, 本文将领域知识划分为两个层次: 一是文本领域知识, 即歧义词所处的句子、段落或文档的领域属性; 二是词义领域知识, 即歧义词各个词义的领域标注属性. McCarthy 等[16]、Jin 等[24]、Gale 等[15] 的相关工作表明, 歧义词在某一特定领域文本中, 其词义的偏斜 (Sense skew) 会非常明显; Magnini 等[8] 的相关工作证明, 充分利用词义的领域信息, 能有效提高词义消歧的正确率. 通过对这两类领域知识的挖掘利用, 改进基于知识库的图模型词义消歧方法在特定领域上的适应性, 也正是本文的主要出发点.

根据 Aitor 等[20]、Stevenson 等[10]、陈文亮等[23] 的相关工作, 文本领域知识可以通过一组文本领域关联词来表示. 文本领域关联词是指与领域文本密切相关、能够表达其主题倾向的词语, 它可作为对歧义词局部上下文的补充. 如 culture 作为名词, 可归纳为两个词义: 1) 文化、文明; 2) (细胞、菌) 培养. 该词可能会出现在社会经济或微生物等领域, 假定社会经济领域关联词为 health、education、income、social 等, 微生物领域关联词为 cell、virus、inhibition、assay 等. 显然, 当 culture 出现在微生物领域时, 其相应的领域关联词对其词义具有明确的指示作用[10].

词义领域知识即词义的领域标注. WordNet Domain 已为 WordNet 的全部同义词集标注了领域信息, 共有 168 个领域标签; 可分为 4 个层次, 其中 Top-level 领域 5 个, Basic 领域 45 个, Factotum 领域 8 个; 不同的 Basic 领域包含若干个 3 级或 4 级领域[8]. 词义领域知识可作为消歧的判定知识. 如名词 fan 有 3 个词义: 1) 风扇; 2) 体育运动爱好者; 3) 狂热追求者. 根据 WordNet Domain, 这 3 个词义的领域标注分别为 factotum、sport、person. 显然, 当 fan 出现在体育领域文本时, 词义 2) 是正确词义的概率将远高于另两个词义.

2.3 基于领域知识的图模型词义消歧技术路线

词义消歧的一个基本原则为 “观其伴、知其义”. 歧义词的词义可根据上下文环境确定. 通常在考虑上下文环境时, 往往仅限于歧义词所在的句子, 提取歧义句中的词形、词性、句法关系等作为消歧特征, 而忽略了领域知识特征. 既然歧义词适配于当前上下文环境, 那么歧义词的正确词义同样应适配于上下文所蕴含的领域环境. 针对文本领域知识和词义领域知识, 本文作以下假设:

假设 1. 相对于其他词义, 歧义词的正确词义与歧义词所属文档的文本领域知识更为适配, 即正确词义与歧义词所属文档的文本领域关联词的关联更加密切.

假设 2. 相对于其他词义, 歧义词的正确词义的词义领域知识与歧义词所属的文档领域更为适配, 即正确词义的词义领域标注与歧义词所属的文档领

域的关联更加密切.

基于这两条假设, 可对图模型进行改进, 提高在特定领域上的适应能力. 本文提出的基于领域知识的图模型词义消歧方法, 首先获取文本领域知识和词义领域知识, 结合句子上下文知识, 根据 BabelNet 创建并调整消歧图; 然后对消歧图进行评分, 选择重要度最高的词义作为正确词义, 整体流程框架如图 1 所示, 具体描述如下: 1) 通过对领域语料库进行对数似然统计, 获得歧义词的文本领域关联词, 作为文本领域知识; 2) 对歧义句进行词形还原后, 获取上下文实词, 作为句子上下文知识; 3) 基于句子上下文知识和文本领域知识, 根据 BabelNet 蕴含的语义关联关系, 构建消歧图; 4) 根据 WordNet Domain 获得歧义词各个词义的领域标注, 作为词义领域知识; 5) 利用词义领域知识并结合歧义词文档领域, 对消歧图进行调整, 生成领域消歧图; 6) 利用图算法对领域消歧图中的词义概念结点进行评分, 获得各概念结点的重要度; 7) 根据各词义概念结点的重要度, 选择重要度最高的词义作为正确词义输出.

2.4 领域知识的获取

2.4.1 文本领域知识的获取

Reddy 等[21]、Stevenson 等[10] 采用对数似然统计的方法, 获取与目标领域有密切关系的领域关键词, 取得了较好的效果. 本文采用对数似然比 (Log likelihood ratio, LLR) 作为词语与领域密切程度的评价指标.

为计算词语 w 与文本领域 D 的对数似然比, 定义如下参数:

1) $a = freq(w, D)$, 包含词语 w, 且属于领域 D 的文档的总数;

2) $b = freq(w, *) - a$, 包含词语 w, 但不属于领域 D 的文档的总数;

3) $c = freq(*, D) - a$, 属于领域 D, 但不含词语 w 的文档的总数;

4) $d = N - a - b - c$, 既不包含词语 w 又不属于领域 D 的文档的总数;

5) N, 语料库包含的全部文档的总数.

对数似然比 LLR 可由式 (1) 计算[25]:

$$\begin{aligned} LLR(w, D) = 2\big[&\log L(p_1, a, a+b) + \\ &\log L(p_2, c, c+d) - \log L(p, a, a+b) - \\ &\log L(p, c, c+d)\big] \end{aligned} \tag{1}$$

其中,

$$\log L(p, k, n) = k\log(p) + (n-k)\log(1-p),$$

$$p_1 = \frac{a}{a+b},\ p_2 = \frac{c}{c+d},\ p = \frac{a+c}{a+b+c+d},$$

$$\log(0) = 0$$

计算出各个词语的 LLR 后, 将词语按 LLR 降序排列; 取其中 top-K 个词语作为文本领域关联词, 作为特定领域 D 的文本领域知识.

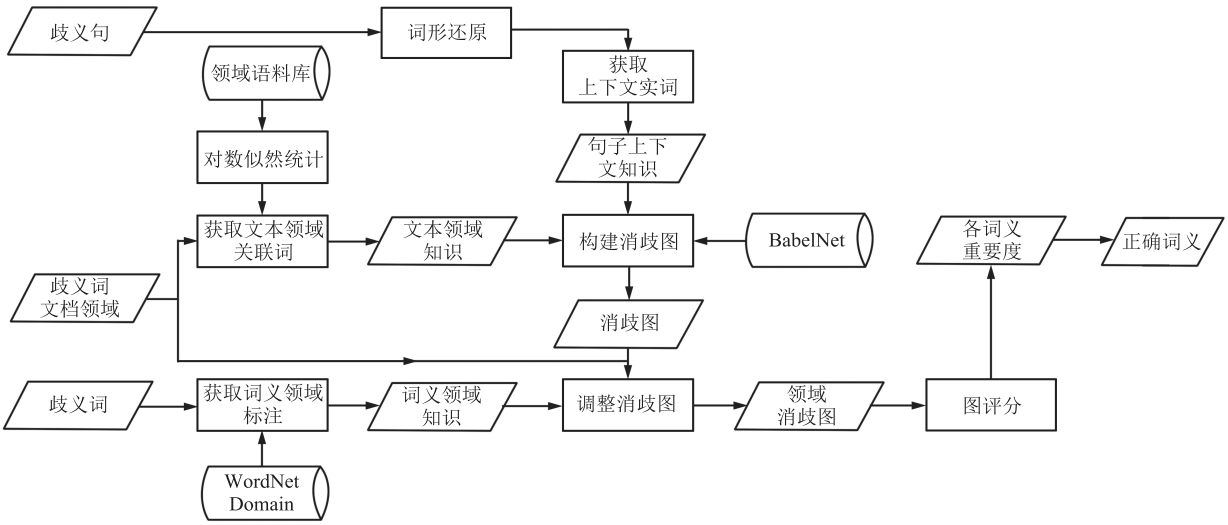


图 1 基于领域知识的图模型词义消歧方法的总体框架

Fig. 1 Framework of WSD method with graph model based on domain knowledge

2.4.2 词义领域知识的获取

WordNet Domain 为 WordNet 中的同义词集逐个进行了领域标注[8]. 对于各个词义概念, 取其对应的 WordNet Domain 的领域标注, 即得其词义领域知识.

2.5 构建消歧图

本文以 BabelNet 作为构建消歧图的依据. BabelNet 包含有 WordNet 的全部同义词集, 并补充了 Wikipedia 蕴含的词义概念, 包含 5.5 M 个概念; 综合了 WordNet 与 Wikipedia 的语义关系, 其中包含 26 种 WordNet 关系、2 种 WordNet Glosses 关系、1 种 Wikipedia 关系[9]. BabelNet 的丰富词义概念及语义关系, 为消歧图的构建提供了很好的支撑.

消歧图的构建需要确定图中的顶点 (词义概念结点) 和边 (词义关联关系). 传统的消歧图的构建, 仅选择句子上下文词语所对应的词义集合作为消歧图中的概念结点. 基于假设 1, 为增强对特定领域的适应能力, 本文将歧义词所属文档领域的文本领域关联词和上下文词语所对应的词义集合, 共同作为消歧图中的概念结点; 利用 BabelNet, 寻找概念结点之间存在的词义关联关系, 作为消歧图中的边.

例如, 对于 Sports 领域文档中的句子 "The coach of the team has left Shanghai by ship", 其中, ship$\#n$ 与 Shanghai$\#n$ 为单义词, coach$\#n$, team$\#n$, leave$\#v$ 为歧义词. 若仅考虑当前句子中的词语作为消歧上下文, 利用 BabelNet 构建其消歧图, 如图 2 所示 (因页面篇幅限制, 图中并未将所有关联路径全部画出. 后续其他消歧图均省略了部分路径). 如果从各个义项结点的出入度来考查, 由图 2 可判断, coach$\#n$ 的义项应为 coach$\#n\#5$ (a vehicle carrying many passengers; used for public transport; "he always rode the bus to work"), leave$\#v$ 的义项应为 leave$\#v\#5$, team$\#n$ 的义项应为 team$\#n\#1$. 显然, coach$\#n$ 的词义判断错误.

假定 Sports 领域的文本领域关联词为 athlete$\#n$, football$\#n$, training$\#n$, 若将文本领域关联词与歧义句词语共同作为消歧上下文, 利用 BabelNet 构建其消歧图, 如图 3 所示 (因消歧图的上半部分没有变化, 与图 2 相同, 故图 3 只画出了受到影响的下半部分). 由图 3 可见, 文本领域关联词的加入, 明显增加了 coach$\#n\#1$ 结点的关联路径数量; 根据结点的出入度可判断, coach$\#n$ 的义项应为 coach$\#n\#1$ ((sports) someone in charge of training an athlete or a team), 修正了单纯依赖句子上下文词语的错误判定.

2.6 调整消歧图

根据 Gale 等[15]、Magnini 等[8]、McCarthy 等[16] 及 Jin 等[24] 的相关研究, 对于特定领域文本, 歧义词的词义偏斜表现更加明显, 更加倾向于选择词义领域与当前文本领域密切相关的词义. 基于假设 2, 本文对词义领域与文本领域密切相关的词义结点的关联关系进行调整.

2.6.1 领域相关度的判定

对于词义领域与文本领域的领域相关度, 本文借助于搜索引擎计数和点式互信息 (Point-wise mutual information, PMI) 来计算. 为了计算词义领域 d_1 与文本领域 d_2 的 PMI, 定义如下参数:

1) $a = freq(d_1, d_2)$, 包含词语 d_1, d_2 关键字的搜索引擎计数;

2) $b = freq(d_1)$, 包含 d_1 关键字的搜索引擎计数;

3) $c = freq(d_2)$, 包含 d_2 关键字的搜索引擎计数;

4) N, 互联网网页的总数 (本文用一个较大的常数 ($N = 10^{11}$) 作为互联网网页总数).

点式互信息 PMI 可由下式计算:

$$\mathrm{PMI}(d_1, d_2) = \frac{N \times a}{b \times c} \tag{2}$$

本文收集 Google 对各个领域关键字的搜索计数, 计算各个领域对的 PMI 值; 取高于一定阈值的领域对作为密切相关的领域对.

2.6.2 消歧图的调整

如果某一词义的词义领域与歧义词所在文档的文本领域被判定为密切相关的, 基于假设 2, 本文认为该词义应为歧义词在当前文本领域的优选词义, 对该词义结点的关联路径进行调整, 以增加其在消歧图中的重要度. 调整算法如算法 1 所示.

算法 1. 消歧图调整算法

输入. 单词 w 的词义集合 $S_w = \{s_i | s_i \in sense(w)\}$; 各个词义 s_i 的领域标注集合 D_{s_i}; 歧义词文档的文本领域标注记作 d_t; 消歧图 G.

输出. 调整后的消歧图 G.

For each s_i in S_w

  Get its D_{s_i}

  For each d_{s_i} in D_{s_i}

    If d_{s_i} 与 d_t 是密切相关的, 且词义结点 s_i 不是词

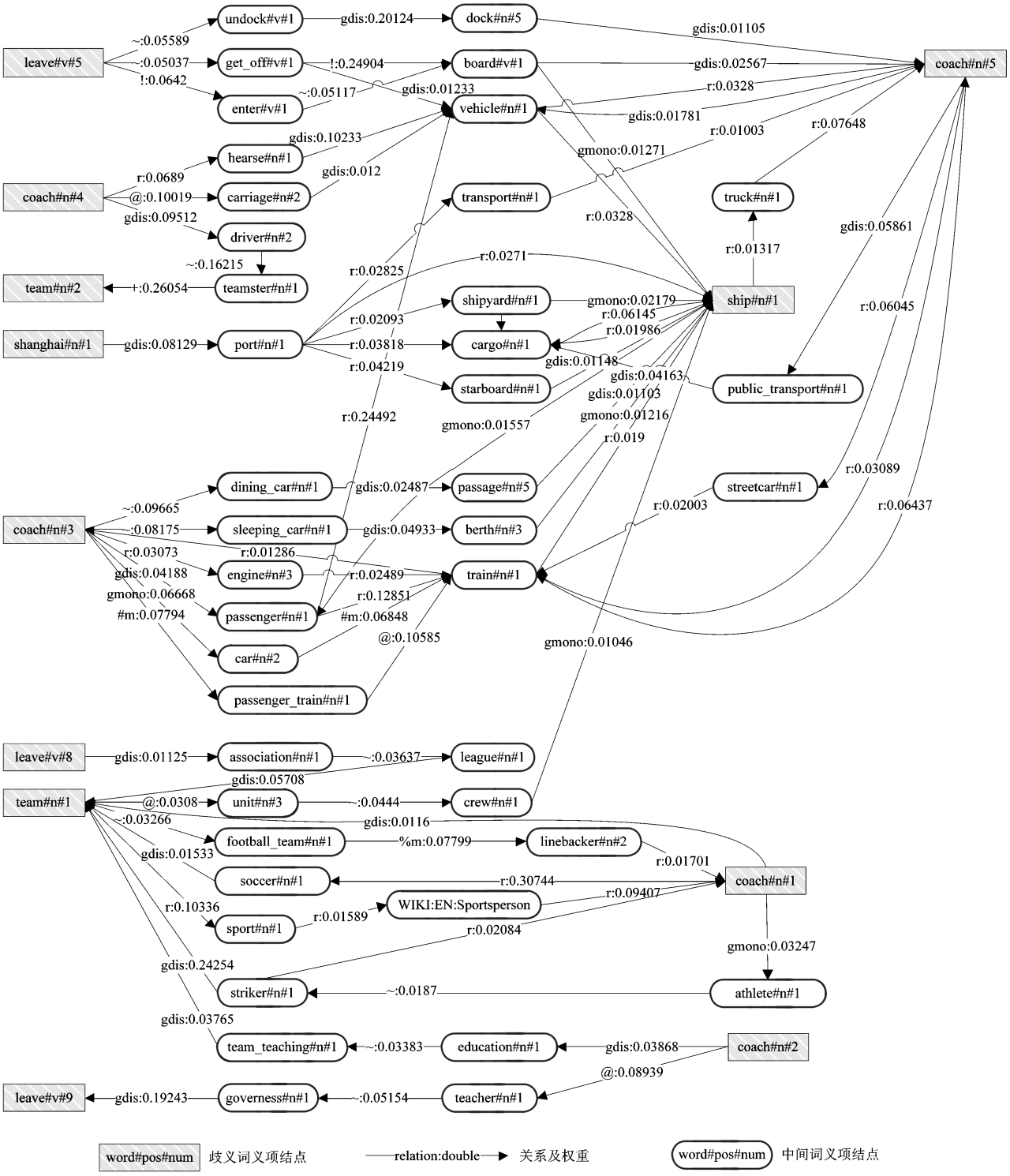


图 2 仅使用歧义句词语构建的消歧图

Fig. 2 Knowledge graph built with context words in ambiguous sentence

义集合 S_w 的度数、关联路径数最大的结点.

1) 根据消歧图 G, 分别取得 S_w 中度数最高的结点 n_d 和关联路径最多的结点 n_p.

2) 增大 s_i 的度数. 根据 n_d 的出度 d^+, 在 G 中添加相应数量的一级虚拟结点; 为 s_i 结点建立到虚拟结点的出弧. 类似的, 根据 n_d 的入度 d^-, 为 s_i 结点构建入弧. 出弧 $outarc$ 和入弧 $inarc$ 的权重根据式 (3) 和式 (4) 设定:

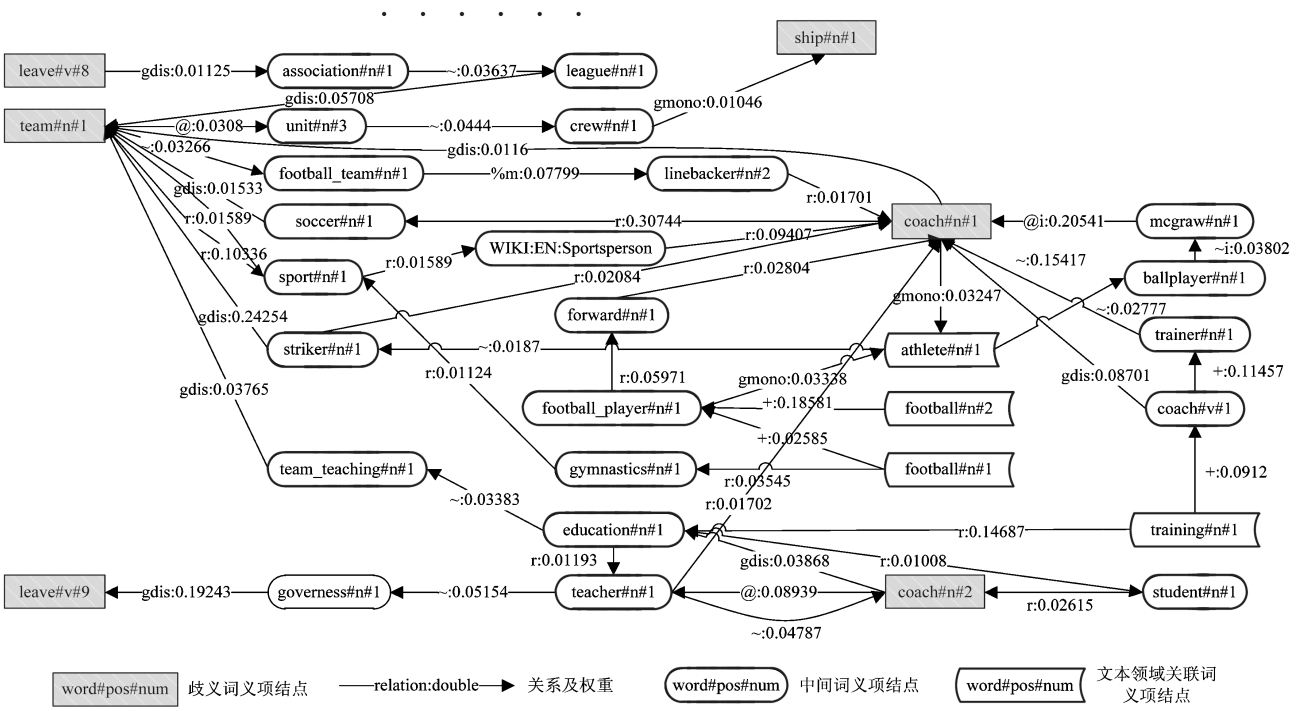


图 3　共同使用文本领域关联词和歧义句词语构建的消歧图 (部分)

Fig. 3　Knowledge graph built on keywords related with text domain and context words in ambiguous sentence (Part)

$$weight_{outarc} = \max_{arc_x \in outarcs(n_d)} (weight(arc_x)) \tag{3}$$

$$weight_{inarc} = \max_{arc_x \in inarcs(n_d)} (weight(arc_x)) \tag{4}$$

其中, $outarcs(n_d)$ 和 $inarcs(n_d)$ 表示 n_d 结点发出的出弧集合和入弧集合.

3) 增加 s_i 的关联路径. 按照 n_p 发出的关联路径数 p^+, 在 G 中添加相应数量的二级虚拟结点; 在 d^+个一级虚拟结点与这些二级虚拟结点之间建立 p^+ 条关联边, 记作 $outedge$. 类似地, 根据进入 n_p 的关联路径数 p^-, 建立相应数量的关联边, 记作 $inedge$. 新建关联边的权重根据式 (5) 和式 (6) 设定:

$$weight_{outedge} = \max_{arc_x \in secondarcs(outpaths(n_p))} (weight(arc_x)) \tag{5}$$

$$weight_{inedge} = \max_{arc_x \in secondarcs(inpaths(n_p))} (weight(arc_x)) \tag{6}$$

其中, $outpaths\,(n_p)$ 和 $inpaths\,(n_p)$ 表示 n_p 结点发出的路径集合和进入的路径集合; $secondarcs(paths)$ 表示路径集合 $paths$ 中的各个路径的第二条边组成的集合.

Break

EndIf

EndFor

EndFor

Return G

该调整方法参照消歧图中度数最高的结点和关联路径数量最多的结点, 为词义领域与文本领域密切相关的词义概念结点人工添加虚拟关联结点和虚拟关联路径, 增加其在消歧图中的重要度. 仍以图 2 为例, 仅根据该消歧知识图, 会误选词义 coach#n#5 作为消歧结果. 根据 WordNet Domain, 可以获得 coach#n 各个词义的领域标注, 其中只有 coach#n#1 的领域标注为 Sport, 其领域标注与歧义词所在的文档领域类型 (Sports) 完全匹配. 图 2 中, 显然 coach#n#5 结点的度数和关联路径数量最多, 其将被选作参照结点. 根据消歧图调整算法, coach#n#1 结点会参照 coach#n#5 结点, 添加出弧、入弧、关联边. 调整后的消歧图中, coach#n#1 结点的度数和关联边的数量均会多于 coach#n#5. 这样便可以保证在对图中结点进行重要度评分时, coach#n#1 的重要度高于 coach#n#5, 促使消歧图输出 coach#n#1 作为正确词义.

2.7　图评分方法

构建消歧图后, 需对图中各概念结点的重要度

进行评分; 为各个歧义词选择评分最高的词义概念作为正确词义. 为便于说明图评分方法, 定义以下符号:

1) s: 词义概念;

2) u, v: 消歧图中的概念结点;

3) V: 消歧图中的概念结点的集合;

4) e: 词关联边;

5) $e_{(u,v)}$: 由概念结点 u 到达结点 v 的一条关联边;

6) $edges_{(u,v)}$: 由概念结点 u 到达结点 v 的关联边集合 (单向);

7) $edges_{(u:v)}$: 概念结点 u 和结点 v 之间的关联边集合 (双向);

8) w_e: 关联边 e 的权重;

9) E: 消歧图中的全部关联边的集合;

10) E^{wiki}: 消歧图中的来自于 Wikipedia 的关联边的集合;

11) p: 两个结点之间的关联路径;

12) $p_{(u,v)}$: 由结点 u 到达结点 v 的最短路径;

13) $paths_{(u:v)}$: 由结点 u 到达结点 v 和由结点 v 到达结点 u 的最短路径的集合.

2.7.1 BabelNet 评分方法

图评分的一个基本依据是: 如果一个结点与其他结点之间存在的关联越多, 则该结点在图中的地位越重要[26]. BabelNet 提供了 Degree、PageRank、Sun-inverse-path-length 和 Sum-path-probability 四种图评分方法[9].

1) Degree

Degree 方法将词义结点的出度作为其重要度, 可表示为

$$score(s) = outdegree(u_s) = |\{e_{(u_s,v)} \in E : v \in V\}| \quad (7)$$

2) PageRank (PR)

PageRank 方法将一个结点的 PageRank (PR) 值作为其重要度. PR 值根据马尔科夫链模型 (Markov chain model) 迭代计算而得. BabelNet 提供的 PageRank 评分方法, 忽略掉了关联边的权重信息, 将消歧图转换为无权有向图, 借助于 Jung-algorithms 工具包, 利用 PageRank 算法对各结点进行评分.

3) Sum-inverse-path-length (SIPL)

Sum-inverse-path-length 方法根据一个结点到其他结点的最短路径的长度倒数之和确定其重要程度, 可表示为

$$score(s) = sipl(u_s) = \sum_{v \in V, v \neq u_s} \frac{1}{\exp(length(p_{(u_s,v)}) - 1)} \quad (8)$$

4) Sum-path-probability (SPP)

Sum-path-probability 方法首先计算一个结点到其他结点的路径上的各条关联边的乘积, 作为路径的重要程度评分; 然后将各条路径的评分之和作为当前结点的重要程度, 可表示为

$$score(s) = spp(u_s) = \sum_{v \in V, v \neq u_s} \prod_{e \in p_{(u_s,v)}} \alpha_e w_e \quad (9)$$

其中,

$$\alpha_e = \begin{cases} \lambda, & \text{若 } e \in E^{wiki} \\ 1 - \lambda, & \text{若 } e \notin E^{wiki} \end{cases}$$

从评分机制是否考虑关联边的权重、是否考虑双向路径来分析, 除 Sum-path-probability 外, 其他 3 类方法均未考虑关联边的权重问题; Degree、Sum-inverse-path-length、Sum-path-probability 均只考虑了由当前结点到达其他结点的路径, 而未考虑由其他结点到达当前结点的路径. 本文认为, BabelNet 的 4 种评分方法均不够完善; 关联边的权重和由其他结点到达当前结点的路径, 对于词义结点重要度的判断具有重要作用.

2.7.2 改进的评分方法

针对 BabelNet 评分方法存在的问题, 本文从关联边和双向路径两个角度, 提出 8 种改进评分方法.

1) Degree-weight (DW)

Degree-weight 方法将词义结点发出的关联边的权重插值累加作为其重要度, 可表示为

$$score(s) = dw(u_s) = \sum_{e \in edges(u_s,v), v \in V, v \neq u_s} \alpha_e w_e \quad (10)$$

其中, α_e 的定义同 Sum-path-probability (α_e 的后续出现均如此定义, 不再单独说明).

2) Degree-outin (DOI)

Degree-outin 方法将词义结点的出度和入度之和作为其重要度, 可表示为

$$score(s) = doi(u_s) = |\{e_{(u_s,v)} \in E \cup e_{(v,u_s)} \in E : v \in V\}| \quad (11)$$

3) Degree-outin-weight (DOIW)

Degree-outin-weight 方法将词义结点的全部关联边的权重插值累加作为其重要度, 可表示为

$$score(s) = doiw(u_s) = \sum_{e\in edges(u_s:v),v\in V,v\neq u_s} \alpha_e w_e \quad (12)$$

4) PageRank-weight (PRW)

PageRank-weight 方法采用带权有向图来评估各个顶点的重要度. 构建带权有向图时, 各个有向边的权重根据关联边的原始权重由式 (13) 确定; 利用 Jung-algorithms 工具包提供的 PageRank 算法对各结点进行评分.

$$weight(e_{(u,v)}) = \frac{\alpha_{e_{(u,v)}} w_{e_{(u,v)}}}{\sum_{e\in edges_{(u,v')},v'\in V,v'\neq u} \alpha_{e_{(u,v')}} w_{e_{(u,v')}}} \quad (13)$$

5) Sum-inverse-path-length-weight (SIPLW)

Sum-inverse-path-length-weight 方法根据一个结点到其他结点的最短路径的长度和该路径上各个关联边的权重插值累加之和, 确定当前结点的重要程度, 可表示为

$$score(s) = siplw(u_s) = \sum_{v\in V,v\neq u_s} \frac{\sum_{e\in p_{(u_s,v)}} \alpha_e w_e}{\exp(length(p_{(u_s,v)}) - 1)} \quad (14)$$

6) Sum-inverse-path-length-outin (SIPLOI)

Sum-inverse-path-length-outin 方法根据由当前结点到达其他结点及由其他结点到达当前结点的最短路径的长度倒数之和, 确定其重要程度, 可表示为

$$score(s) = siploi(u_s) = \sum_{v\in V,v\neq u_s} \sum_{p\in paths_{(u_s:v)}} \frac{1}{\exp(length(p) - 1)} \quad (15)$$

7) Sum-inverse-path-length-outin-weight (SIPLOIW)

Sum-inverse-path-length-outin-weight 方法根据由当前结点到达其他结点及由其他结点到达当前结点的最短路径的长度和最短路径上的关联边的权重插值累加之和, 确定其重要程度, 可表示为

$$score(s) = siploiw(u_s) = \sum_{v\in V,v\neq u_s} \sum_{p\in paths_{(u_s:v)}} \frac{\sum_{e\in p} \alpha_e w_e}{\exp(length(p) - 1)} \quad (16)$$

8) Sum-path-probability-outin (SPPOI)

Sum-path-probability-outin 方法针对由当前结点到达其他结点及由其他结点到达当前结点的最短路径的集合, 计算各条路径上的关联边的权重乘积, 作为该路径的重要度评分; 而后将全部路径的评分之和作为当前结点的重要度, 可表示为

$$score(s) = sppoi(u_s) = \sum_{v\in V,v\neq u_s} \sum_{p\in paths_{(u_s:v)}} \prod_{e\in p} \alpha_e w_e \quad (17)$$

3 实验

3.1 数据集和评价指标

基于领域知识的图模型词义消歧方法, 是针对特定领域的词义消歧任务而提出的. 目前最常用的特定领域词义消歧任务的实验数据集为 Koeling 数据集[14]. 为便于与相关研究进行比较, 本文采用该数据集作为实验数据. Koeling 数据集是一个采样词任务 (Lexical sample task) 数据集, 消歧实例分别来自 BNC (British National Corpus) 语料库及 Reuter 语料库的 Sports 和 Finance 领域语料库, 共包含 41 个采样词, 采样词的义项数范围为 2∼13, 平均义项数为 6.7. 该数据集中的每个消歧实例均由 2∼3 人进行人工标注, 取得多数标注一致的实例共 9 568 个.

本文采用召回率 (Recall) 作为评价指标. 设 A 表示待消歧实例个数, B 表示得到正确消歧结果的实例个数, 则召回率 R 可由公式 $R = B/A$ 计算.

3.2 实验参数设定

本文实验涉及的参数较多, 下面分别予以说明:

1) 文本领域关联词的获取语料

本文利用 Reuter 语料库为 Sports 领域和 Finance 领域获取文本领域关联词. Reuter 语料库中的所有文档均已进行了人工主题分类, 各文档均标有主题分类代码. 本文选择将主题代码 (Topic codes) 包含 ECAT 或 MCAT 的 117 734 篇文档构成 Finance 领域语料库, 将主题代码包含 GSPO 的 35 317 篇文档构成 Sports 领域语料库, 将全体 Reuter 语料库作为整体语料库.

2) 文本领域关联词的数量

根据对数似然比, 本文为 Sports 领域和 Finance 领域获得与其最为密切的关联词语列表. 在实验中各取前 10 个词语作为相应领域的领域关联词.

3) 领域相关度的判定阈值

本文将领域相关度的判定阈值设为 1.48. 若词义领域与文本领域的 PMI 值高于该阈值, 则认为两者密切相关; 根据假设 2, 相应结点的关联边将被调整扩充.

4) BabelNet 路径过滤参数设定

本文将路径过滤参数 knowledge.graph.filters 设为 Loop、Illegal-pointers、Sense-shifts、Min-weight. 当创建消歧图时, 凡是满足以上过滤条件的路径均会被剔除. 上述各参数解释如下:

a) Loop: 检查一条路径上是否存在环.

b) Illegal-pointers: 检查一条路径上的各条边的语义关系类型是否为不适当的. 本文使用工具包缺省设置, 置 knowledge.graph.filter.illegalpointers 为 "; c,; r,; u, -c, -r, -u".

c) Sense-shifts: 检查一条路径中是否包含了同一个单词的不同词义结点.

d) Min-weight: 检查一条路径中的各条边的权重是否低于指定的最小值. 本文使用工具包缺省设置, 置 knowledge.graph.filter.weight.threshold 为 0.01.

3.3 实验结果

3.3.1 不同评分方法的效果对比实验

在第 2.7 节中, 除了 BabelNet 自带的 4 种评分方法, 又提出了 8 种改进评分方法. 本节首先进行一组实验, 比较这 12 种评分方法的性能优劣, 并寻找最佳的关联边权重插值系数. 实验结果如图 4 所示.

由图 4 的实验数据可以看出:

1) SIPLOI、DOI、PR、SIPL、Degree 5 类方法的效果始终保持不变; 这 5 类评分方法并未考虑关联边的权重问题, 其性能不受权重插值系数的影响.

2) 随着插值系数的增大, SIPLOIW、DOIW、SPPOI、PRW、SIPLW、SPP、DW 7 类方法的性能均表现出先增后减的变化趋势. 这说明单独依赖于 Wikipedia 关系或 WordNet 关系均无法取得最优消歧效果.

3) 在各类评分方法中, 当 $\lambda = 0.3$ 时, SIPLOIW 评分方法取得最佳消歧效果, 召回率达到 54.79 %; DOIW 评分方法取得次优消歧效果, 召回率达到 54.71 %; $\lambda = 0.3$, 意味着在对关联边评分时, 来自于 Wikipedia 的关联边被弱化了. 当 $\lambda = 0$ 时, 意味着来自于 Wikipedia 的关联边的权重被忽略, 只考虑来自于 WordNet 的关联边权重; $\lambda = 1$, 则反之; 对比 $\lambda = 0$ 和 $\lambda = 1$ 可发现, 前者的性能普遍优于后者. 这两方面均说明: Wikipedia 的关联边的质量要次于 WordNet 的质量. Wikipedia 中存在的主要关系为不同概念之间的超链接关系,这相比于

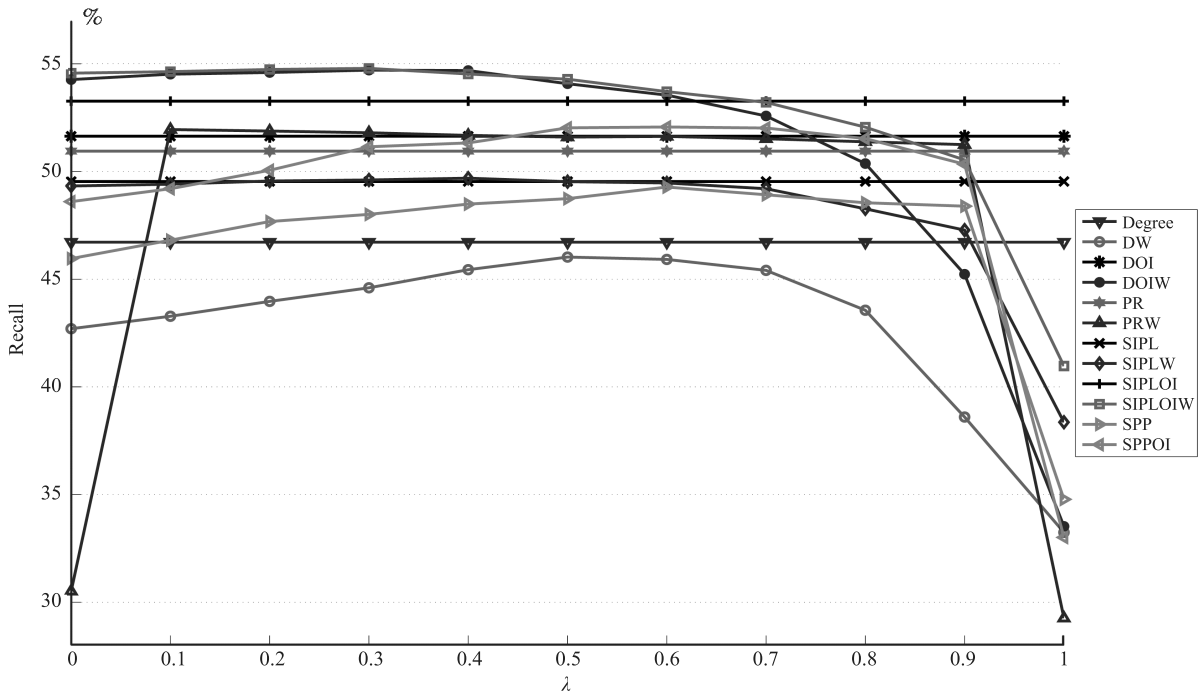


图 4 不同插值系数下各类评分方法性能对比

Fig. 4 Effectiveness comparison of evaluation methods with different interpolation parameters

WordNet 中上下位、整体部分等语义关系的准确性要差, 理应相对弱化一些.

4) 将 Degree、DW、DOI、DOIW 对比, 除了 DW 方法, 改进的 DOI、DOIW 的效果整体要远远优于标准的 Degree 方法. 将 PR、PRW 对比, PRW 整体优于标准 PR 方法. 将 SIPL、SIPLW、SIPLOI、SIPLOIW 对比, 其中 SIPLW 略优于标准 SIPL、SIPLOI 和 SIPLOIW 远远优于标准的 SIPL 方法. 将 SPP 与 SPPOI 对比, 后者明显优于前者. 总体来看, 各类改进评分方法的效果要优于原标准方法. 这说明本文提出的通过考虑关联边的权重和双向路径来改进原标准评分方法是切实有效的, 可明显改善消歧效果.

5) 单独考查 BabelNet 自带的 Degree、PR、SIPL 和 SPP 4 种标准评分方法, PR 方法的效果最优, 可达 50.95 %. 但综合考查所有评分方法, 计算复杂度最高的 PageRank 类算法 (包括 PR、PRW), 其能取得最优效果仅为 51.95 % (PRW, $\lambda = 0.1$), 明显低于 SIPLOIW、DOIW 和 SIPLOI 方法. 在词义消歧中, 复杂度高的 PageRank 类图评分算法未必会有最佳效果, 相反, 越简单的算法等却相对更加有效. 这与 Navigli 等的相关研究[26] 结论是一致的.

3.3.2 本文方法与其他文献方法的效果对比

鉴于本文提出的评分方法中, SIPLOIW 和 DOIW 在 $\lambda = 0.3$ 时取得了较好的消歧效果, 本文将这两种评分方法作为本文方法的代表; 将 PR 作为 BabelNet 自带标准评分机制的代表. 将它们与其他文献中的方法进行横向对比, 如表 1 所示.

表 1 本文方法与其他文献方法的消歧效果对比 (%)

Table 1 Effectiveness comparison of our method and other methods (%)

Systems	BNC	Sports	Finance
MFS	34.9	19.6	37.1
SVM	38.7	25.3	38.7
Koeling	40.7	43.3	49.7
PPRank-context	43.8	35.6	46.9
PPRank-related Words	37.7	51.5	59.3
PR	40.83	54.95	57.19
SIPLOIW	**45.58**	55.9	62.94
DOIW	43.53	**56.92**	**63.78**

对比的消歧方法包括 MFS、SVM、Koeling、PPRank-context 和 PPRank-related Words, 简要介绍如下:

1) MFS: 根据 SemCor 统计数据, 选择各个单词的最大词频词义 (Most frequent sense, MFS), 相当于根据 WordNet 词频选择最常用的词义.

2) SVM: 由 Agirret 等实现的一种有监督的词义消歧方法[7]. 该方法使用局部搭配、句法依存关系及词袋作为消歧特征, 并在 SemCor 语料上训练 SVM 分类器.

3) Koeling: 实验数据集的开发者 Koeling 等提出的一种自动获取特定领域优势词义的消歧方法[7, 14]. 该方法首先为歧义词获取在特定领域语料上依存分布最为相似的 50 个词语, 然后计算它们与歧义词各义项的相似度, 选择相似度之和最大的义项作为消歧结果.

4) PPRank-context: 该方法使用 Agirre 等提出的 Personalized PageRank 算法, 使用当前句子中的词语作为消歧上下文, 判断正确词义[7].

5) PPRank-related Words: 该方法利用 Koeling 等提出的方法为歧义词获取在特定领域上依存分布最为相似的 50 个词语作为消歧上下文[14], 使用 Agirre 等提出的 Personalized PageRank 算法判断正确词义[7].

由表 1 的实验结果可以看出:

1) MFS 和 SVM 方法的效果明显次于其他方法. 由于歧义词的词义在不同领域上存在较强的词义偏斜问题, 在 SemCor 语料库上统计的 MFS、训练的 SVM 模型均无法有效判断特定领域的歧义词词义.

2) Koeling 与 PPR-related Words 均使用依存分布最为相似的 50 个词语作为消歧上下文. 后者在特定领域 (Sports, Finance) 上的效果明显优于前者, 说明图模型方法 (PPR-related Words) 优于相似度度量 (Koeling) 的方法.

3) PPRank-context 使用当前句子中的词语作为消歧上下文, 在特定领域上, 其效果次于 Koeling 和 PPR-related Words 方法. 这说明对于特定领域词义消歧, 单独依靠歧义句所包含的词语作为消歧线索并不充足, 必须为其收集一定数量的领域关联词作为补充. 这也从侧面验证了本文获取领域知识改善特定领域词义消歧效果这一出发点的可行性和必要性.

4) 对比 BNC 数据的实验结果, SIPLOIW 方法最佳, 其次为 PPRank-context, 之后为 DOIW. 对

于 BNC 数据, 因为其无明确领域信息, 本文未补充任何文本领域关联词, 也未对消歧图作任何干涉调整; 然而, 基于 BabelNet 知识库的 SIPLOIW 方法和 DOIW 方法仍分别取得了第 1 名和第 3 名. 证明整合了 WordNet 和 Wikipedia 的 BabelNet 知识库的有效性; 为基于 BabelNet 知识库的方法取得更优效果提供了有利支撑条件.

5) 对比 Sports 和 Finance 领域数据的实验结果, 本文提出的 DOIW 和 SIPLOIW 方法均明显优于其他文献的方法, 充分证明了本文方法的有效性. 本文提出的以 BabelNet 为知识库, 结合文本领域知识和词义领域知识构建并调整消歧图, 对图中概念结点进行评分以获得正确词义的消歧方法是切实可行的.

3.3.3 两个假设的验证实验

第 2.3 节提出了两个假设. 基于这两个假设, 本文取文本领域关联词作为文本领域知识, 与歧义句所包含的上下文词, 共同构建消歧图; 根据词义领域标注与歧义词文档领域的相关性, 对消歧图进行调整. 为了验证这两个假设的正确性, 本文分别构建另 3 个消歧模型:

1) NoTextDomainWords (NTDW): 该消歧模型在构建消歧图时, 将不再使用文本领域关联词 (文本领域知识); 仅利用歧义句的上下文实词来作消歧上下文.

2) NoSenseDomainAdjust (NSDA): 该消歧模型将不再考虑歧义词的词义领域标注 (词义领域知识) 与文档领域的相关性, 不再对消歧图进行干涉调整.

3) NoTextDomainWordsSenseDomainAdjust (ND): 该消歧模型综合了前两种模型, 既不再利用文本领域知识, 也不再使用词义领域知识.

鉴于在前期实验中, SIPLOIW 和 DOIW 方法取得了最佳的消歧效果. 本文将这两种方法应用于这 3 种模型, 验证两个假设的正确性, 并评价文本领域知识和词义领域知识所发挥的作用. 实验数据分别如图 5 和图 6 所示.

由图 5 和图 6 可以看出, 从 Koeling 数据集的全部消歧实例 (All) 观察数据, 发现 SIPLOIW 和 DOIW 方法的 3 个消歧模型的效果均明显下降; 不使用任何领域知识的 ND 模型的效果下降达 10 % 左右. 从 3 个子数据集 (BNC、Sports、Finance) 观察数据, BNC 数据集的消歧实例未使用任何领域知识, 效果不受消歧模型的影响; Sports 数据集上, SIPLOIW 在 NTDW 模型中意外表现出效果略微提高的现象, DOIW 的表现与 All 数据集相一致; Finance 数据集的表现与 All 数据集完全一致.

总体来看, 这 3 种舍弃领域知识的消歧模型的消歧效果次于本文提出的标准模型的效果. 这也证明了本文提出的两个假设的正确性.

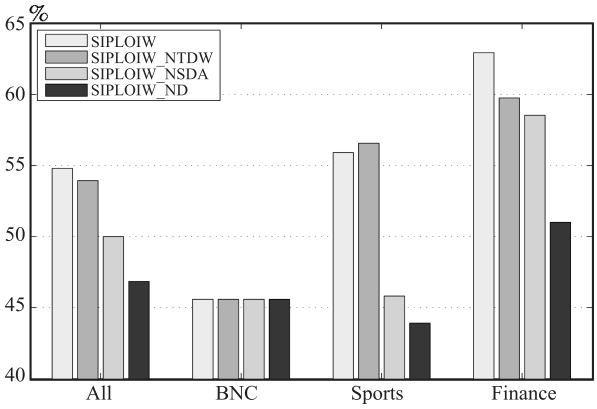


图 5 采用不同消歧模型的 SIPLOIW 方法的效果对比

Fig. 5 Effectiveness comparison of SIPLOIW method with different models

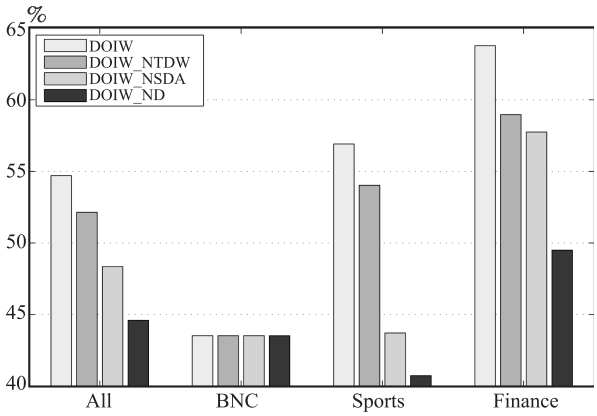


图 6 采用不同消歧模型的 DOIW 方法的效果对比

Fig. 6 Effectiveness comparison of DOIW method with different models

4 总结

本文提出了一种基于领域知识的图模型词义消歧方法. 该方法充分发挥领域知识和图模型的优势, 针对特定领域的消歧实例, 首先利用对数似然统计的方法为目标领域收集文本领域关联词, 作为文本领域知识; 利用文本领域关联词与歧义词的句子上下文, 共同构建消歧图; 而后, 依据 WordNet Domain 为歧义词的各个词义获取词义领域标注, 作为词义领域知识; 根据词义领域标注与歧义词文档领域的密切程度, 对消歧图进行调整; 最后, 利用改进的图评分方法, 对消歧图中的各个词义结点的重要

度进行评分, 选择评分最高的结点作为正确词义输出. 该方法有效地将领域知识整合到图模型中, 深入挖掘图模型的消歧潜力, 在 Koeling 数据集上, 取得了同类研究中最佳的消歧效果. 此外, 本文对多种图模型评分方法进行改进, 并作了详细对比研究, 有效改善了已有评分方法的消歧效果.

本文首次提出分别利用文本领域知识和词义领域知识构建并调整消歧图的图模型词义消歧方法. 为面向特定领域的词义消歧对于领域知识的挖掘利用提供了一个可行的解决方案. 本文对图模型评分方法的对比研究对于同类研究工作亦将具有一定的借鉴价值. 我们下一步的工作将从以下几个方面进行: 1) 考察文本领域关联词和句子上下文实词的消歧权重问题, 文本领域关联词与目标领域的密切程度、上下文实词与歧义词之间的距离理应对消歧效果有所影响; 2) 尝试改进 BabelNet 包含的词义关联关系的质量, 由实验数据来看, BabelNet 包含的 Wikipedia 的词义关联的质量并不高; 3) 发掘更有效的图模型评分机制, 进一步挖掘图模型的消歧潜力; 4) 探索更有效寻找文本领域关联词的方法, 从第 3.3.3 节的实验数据来看, SIPLOIW 在 NTDW 模型中效果反而略有提高, 可能是因 Sports 的领域关联词精度不高所致.

References

1 Navigli R. Word sense disambiguation: a survey. *ACM Computing Surveys*, 2009, **41**(2): 1011−1069

2 Liu Yu-Peng, Li Sheng, Zhao Tie-Jun. System combination based on WSD using WordNet. *Acta Automatica Sinica*, 2010, **36**(11): 1575−1580
(刘宇鹏, 李生, 赵铁军. 基于 WordNet 词义消歧的系统融合. 自动化学报, 2010, **36**(11): 1575−1580)

3 Lu Zhi-Mao, Liu Ting, Li Sheng. The research progress of statistical word sense disambiguation. *Acta Electronica Sinica*, 2006, **34**(2): 333−343
(卢志茂, 刘挺, 李生. 统计词义消歧的研究进展. 电子学报, 2006, **34**(2): 333−343)

4 Wang Bo, Yang Mu-Yun, Li Sheng, Zhao Tie-Jun. Evaluation of all-words WSD for Chinese in machine translation. *Acta Automatica Sinica*, 2008, **34**(5): 535−541
(王博, 杨沐昀, 李生, 赵铁军. 中文全词消歧在机器翻译系统中的性能评测. 自动化学报, 2008, **34**(5): 535−541)

5 Wang Rui-Qin, Kong Fan-Sheng. Research on unsupervised word sense disambiguation. *Journal of Software*, 2009, **20**(8): 2138−2152
(王瑞琴, 孔繁胜. 无监督词义消歧研究. 软件学报, 2009, **20**(8): 2138−2152)

6 Lu Zhi-Mao, Liu Ting, Li Sheng. Full-words automatic word sense tagging based on unsupervised learning algorithm. *Acta Automatica Sinica*, 2006, **32**(2): 228−236
(卢志茂, 刘挺, 李生. 基于无指导机器学习的全文词义自动标注方法. 自动化学报, 2006, **32**(2): 228−236)

7 Agirre E, de Lacalle O L, Soroa A. Knowledge-based WSD and specific domains: performing better than generic supervised WSD. In: Proceedings of the 2009 International Joint Conference on Artificial Intelligence 2009. Pasadena, USA: Morgan Kaufmann Publishers Inc, 2009. 1501−1506

8 Magnini B, Strapparava C, Pezzulo G, Gliozzo A. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 2002, **8**(4): 359−373

9 Navigli R, Ponzetto S P. BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 2012, **193**: 217−250

10 Stevenson M, Agirre E, Soroa A. Exploiting domain information for word sense disambiguation of medical documents. *Journal of the American Medical Informatics Association*, 2011, **19**(2): 235−240

11 Agirre E, de Lacalle O L, Fellbaum C, Hsieh S K, Tesconi M, Monachini M, Vossen P, Seqers R. SemEval-2010 task 17: all-words word sense disambiguation on a specific domain. In: Proceedings of the 2009 NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Boulder, Colorado: Association for Computational Linguistics, 2009. 123−128

12 Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the ACL. Stroudsburg: Association for Computational Linguistics, 2009. 33−41

13 Mihalcea R, Tarau P, Figa E. PageRank on semantic networks, with application to word sense disambiguation. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). Stroudsburg: Association for Computational Linguistics, 2004. Article no. 1126, DOI: 10.3115/1220355.1220517

14 Koeling R, Macarthy D, Carroll J. Domain-specific sense distributions and predominant sense acquisition. In: Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). Stroudsburg: Association for Computational Linguistics, 2005. 419−426

15 Gale W A, Church K W, Yarowsky D. One sense per discourse. In: Proceedings of the 4th DARPA Workshop on Speech and Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 1992. 233−237

16 McCarthy D, Koeling R, Weeds J, Carroll J. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 2007, **33**(4): 553−590

17 Agirre E, de Lacalle O L. Supervised domain adaption for WSD. In: Proceedings of the 12th Conference of the European Chapter of the ACL. Athens, Greece: Association for Computational Linguistics, 2009. 42−50

18 Chan Y S, Ng H T. Domain adaptation with active learning for word sense disambiguation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, 2007. 49−56

19 Zhong Z, Ng H T, Chan Y S. Word sense disambiguation using OntoNotes: an empirical study. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stoudsburg, PA: Association for Computational Linguistics, 2008. 1002−1010

20 Aitor S, Eneko A, Oier L L, Monica M, Jessie L, Shu K H. Kyoto: an integrated system for specific domain WSD. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: Association for Computational Linguistics, 2010. 417−420

21 Reddy S, Inumella A, McCarthy D, Stevenson M. IIITH: domain specific word sense disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation. Stoudsburg, PA: Association for Computational Linguistics, 2010. 387−391

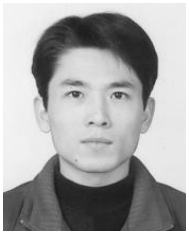
22 Galley M, McKeown K. Improving word sense disambiguation in lexical chaining. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003). San Francisco, CA: Morgan Kaufmann Publishers Inc., 2003. 1486−1488

23 Chen Wen-Liang, Zhu Jing-Bo, Zhu Mu-Hua, Yao Tian-Shun. Text representation using domain dictionary. *Journal of Computer Research and Development*, 2005, **42**(12): 2155−2160
(陈文亮, 朱靖波, 朱慕华, 姚天顺. 基于领域词典的文本特征表示. 计算机研究与发展, 2005, **42**(12): 2155−2160)

24 Jin P, McCarthy D, Koeling R, Carroll J. Estimating and exploiting the entropy of sense distributions. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado: Association for Computational Linguistics, 2009. 233−236

25 Liu Peng-Yuan, Zhao Tie-Jun. Unsupervised translation disambiguation based on Web indirect association of bilingual word. *Journal of Software*, 2010, **21**(4): 575−585
(刘鹏远, 赵铁军. 基于双语词汇 Web 间接关联的无指导译文消歧. 软件学报, 2010, **21**(4): 575−585)

26 Navigli R, Lapata M. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(4): 678−692



鹿文鹏 北京理工大学计算机学院博士研究生, 齐鲁工业大学理学院副教授. 主要研究方向为词义消歧. 本文通信作者. E-mail: luwpeng@bit.edu.cn
(**LU Wen-Peng** Ph. D. candidate at the School of Computer Science and Technology, Beijing Institute of Technology and associate professor at the School of Science, Qilu University of Technology. His main research interest is word sense disambiguation. Corresponding author of this paper.)

黄河燕 北京理工大学教授. 主要研究方向为自然语言处理, 机器翻译. E-mail: hhy63@bit.edu.cn
(**HUANG He-Yan** Professor at Beijing Institute of Technology. Her research interest covers natural language processing and machine translation.)

吴 昊 北京理工大学计算机学院博士研究生. 主要研究方向为语义相似度计算. E-mail: wuhao123@bit.edu.cn
(**WU Hao** Ph. D. candidate at the School of Computer Science and Technology, Beijing Institute of Technology. His main research interest is semantic similarity computation.)