学术文献引文上下文自动识别研究*

雷声伟[1] 陈海华[1] 黄永[1] 陆伟[1,2]
[1]武汉大学信息管理学院 武汉 430072 [2]武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要: [目的/意义] 引文内容分析能够帮助揭示文献引用关系的深层语义内涵,而引文上下文识别作为引文内容分析的基础显得尤为重要。[方法/过程] 梳理已有引文上下文研究的现状,总结当前引文上下文识别的不足,在此基础上归纳引文上下文识别的 5 类特征,并采用文本分类和序列标注两种方法开展引文上下文自动识别实验。[结果/结论] 实验结果表明,本文提出的特征能够很好地提升引文上下文识别效果,且基于文本分类的 SVM 分类效果要优于基于序列标注的 CRF。

关键词: 引文上下文 引文内容分析 支持向量机 条件随机场 隐式上下文
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2016.17.012

1 引言

引文分析一直以来是文献计量学中的重要研究方法,文献的引证数据常常被用于学科的研究热点和趋势分析,但基于被引频次等统计数据的计量分析将所有被引等同起来,无法反映出作者的引用动机、情感倾向等引证行为[1],这些信息都包含在引文内容中,而恰恰是这些内容信息决定了文献被引的价值。因此,将引文的内容信息应用到引文网络分析中,能够为引文网络引入更为丰富可靠的特征,从而使得现有的科学评价和分析更加有效。

当一篇文献引用另一篇文献时,其引文标记所在的句子会包含对该文献的叙述或评论,这种包含特定引文标记的句子称为引文句(citation sentence)或显式引文上下文[2]。引文句是引用的载体,是引文内容分析的重要部分,其标记具有较为统一的格式,很容易通过正则表达式等方法进行识别。引文句和被引文献关系最为密切,通过引文句可以获得很多关于被引文献的信息。然而,仅仅通过引文句来进行引文内容分析存在一定问题: 研究者在引用一篇文献时,可能需要多句话才能完整表达对这篇文献的观点,这些话语同样包含作者对这篇文献的描述和评论,对于引文内容分析也有十分重要的作用,这种类型的上下文称为隐式引文上下文。只有同时获得了包含显式引文上下文和隐式引文上下文在内的完整引文上下文信息,才能准确把握作者对该文献的评价和描述,更准确地进行引文内容分析。目前,引文上下文的识别和相关应用已成为研究热点,引文上下文作为引文内容分析的基础,在任何基于引文内容的研究中都是必不可少的。

由于缺少明显的标记,对隐式引文上下文范围的识别十分困难(如未说明,本文后面提及引文上下文自动识别均指识别隐式引文上下文)。为更准确完整地挖掘出文献中和引文相关的内容,需要结合有效的特征和算法从学术文献中提取完整的引文上下文。

2 相关研究

引文在科研文献中十分普遍,体现了后来研究者对前人研究成果的借鉴和认可,为文献和文献之间建立了一条知识传递的纽带[3]。传统的引文分析通过文献之间的引用关系及相关统计来度量学术成果的影响力,如影响因子、h-index 等。然而,研究者引用文献的动机复杂且多样,有些只是将其作为相关研究、历史背景或未来工作等,有些引用态度甚至是负面的,而将这些不同的引用动机都同等看待显然是不合理的。随

* 本文系国家自然科学基金面上项目"面向词汇功能的学术文本语义识别与知识图谱构建"(项目编号: 71473183)研究成果之一。
作者简介: 雷声伟(ORCID: 0000-0002-7152-7817),硕士研究生; 陈海华(ORCID: 0000-0003-2806-3938),硕士研究生; 黄永(ORCID: 0000-0003-4808-6491),博士研究生; 陆伟(ORCID: 0000-0002-0929-7416),武汉大学信息管理学院副院长,教授,博士生导师,通讯作者,E-mail: weilu@whu.edu.cn。
收稿日期: 2016-06-16 修回日期: 2016-08-15 本文起止页码: 78-87 本文责任编辑: 刘远颖

着研究的深入,一些学者开始质疑传统基于计量的引文分析的可靠性[4-5],开始探讨更为深入的引文评价方法和技术,其中结合引文内容分析是一个较为可行的方向[6]。引文内容分析,就是以引文上下文为依据对引文的性质进行分析,它根据引文内容的语法和语义特征,将被引文献对施引文献支持的作用和程度进行区分,深入挖掘施引文献与被引文献之间的语义关联,进而揭示引用行为的本质[3]。引文内容分析的基础是从文献中识别引文上下文,因此能否正确有效地进行引文上下文识别显得尤为重要。

针对引文上下文识别,已有部分国外学者进行过较长时间的研究。H. Nanba 等[7]使用引用区域(citing area)这一概念来定义引文上下文,引用区域是指引用句周围的一个连续区域。他们首先将包含目标引用的句子作为引用区域的第一个句子,然后使用一组线索词(包括一些指代词语、连接词语和第三人称代词等)和手工指定的规则来判断它周围的句子是否属于该引用区域,并将引用区域的识别范围定在当前段落。他们在人工构造的数据集上进行了引文区域自动识别实验,取得了80%的准确率和76%的召回率,并将该结果用于检索和引文网络中,同时还开发了原型系统用于特定领域文献综述的自动生成。而 A. Abu-Jbara 等[8]则注意到引文句中可能包含多个引用而导致生成的摘要不连贯,提出将与目标引文不相关的片段进行移除可以改善自动摘要的效果,通过生成句法树保留了和目标引文相关的最小数量的片段,但这种方式过于简单,很多情况下并不准确。A. Athar[9]采用类似于 A. Abu-Jbara 等[8]的方式解决多个引用问题,不过他定义了一个引文影响范围(scope of citation influence)来改进分类效果。M. A. Angrosh 等[10]针对文献中"相关工作"章节的引用进行了上下文识别,他们认为引用主要集中在研究进展部分且该处的引用动机很明确,并分析了该章节的一般引用模式,然后基于模板找出上下文的一些特征(如背景、主题、优点、缺点、结果、方法等关键词特征),使用 CRF(条件随机场)分类器进行训练,取得了96.51%的准确率,然而这种方法的应用场景过于局限。事实上,出现在"相关研究"章节的引文上下文往往不能反映出被引文献的核心内容和主要功能,因此该学者所做的研究不是基于引文上下文相关研究领域(如自动摘要、引文推荐、引文功能识别、引文重要性识别等)的重点。另外,不同学科、不同类型文献的"相关研究"撰写也具有差异性,基于模板进行分析也过于僵化。

2010 年,V. Qazvinian 等[11]对引文句和上下文进行了区分,他们将引文句定义为引文标记所在的句子,上下文为非明确描述引文的句子集合,并使用概率模型去抽取隐式的上下文,如果一个句子在引文句周围则建立关联。他们给文章中每个句子均设立一个标签,用来表示该句子是否为上下文,进一步使用 CRF 来进行引文上下文的识别。在多组对照实验中,当候选上下文为前后各 4 个句子时效果最好,但也只在 1 篇文献上的识别效果达到了 88.9% 的 F3 值,10 篇测试文献的平均效果仅为 54.0% 的 F3 值,说明识别的引文上下文中包含许多冗余信息,且不具有通用性,很难付诸实际应用。M. Y. Kan[12]主要通过机器学习方法分别结合 ME 模型和 SVM 模型进行引文句的识别,以词汇特征为主要的分类特征,两种模型的对比结果显示,该分类任务的准确度与分类器的选择没有明显关系。

2012 年,A. Abu-Jbara 等[13]在其之前研究基础上,又使用了 3 种方式进行引文上下文识别:①将句子转换成一个个单词,再通过分类器确定每一个单词是否属于目标引文的上下文;②将其转换成一个序列标注问题并为每个单词分配类别;③将句子切分为不同片段,然后根据片段信息来判断引文句的范围。实验结果表明基于片段分类的效果最好,准确率为 81.8%。随后,他们对第二种方式进行了深入研究,通过 CRF 算法在引文句周围一定窗口(单词跨度)内寻找最优的类别序列,类别包括 include 和 exclude 两类,选择引文句前面一句和后面四句作为候选上下文并应用于引文情感识别,比仅使用引文句作为上下文的效果在 F1 上提升了 12.1%,表明隐式引文上下文包含十分丰富的信息[2]。

2013 年,M. A. Angrosh 等[14]同样使用词汇特征并构造 CRF 模型进行引文上下文识别,开发了引文上下文自动抽取系统 CitContExt;2014 年,P. Sondhi 等[15]构造文献句子个数-引文个数矩阵并使用 HMM(hidden Markov mode)模型进行隐式引文上下文的抽取;2014 年,A. Athar 等[16]结合词汇特征和句法特征并使用 SVM(support vector machine)分类器来进行引文上下文的识别,他们在实验中证明引文上下文对引文情感识别的效果相对提升 48%(F 值),对引文重要性的识别效果相对提升 17%(F 值)。

国内对引文上下文的研究较少,专门进行引文上下文探测方面还属于空白,大多停留在论证引文上下文是否具有对文献内容分析的研究价值上。刘盛博等[17]将引文上下文引入到引文的评价中;刘洋等[1]探

讨了引文上下文在文献内容分析中的信息价值; 许德山[18]利用引文上下文信息进行引用的观点倾向性识别; 孙枫军[19]则直接将引文上下文的范围缩小至引文句,并用于概念抽取; 张金松[20]利用引文上下文的语义信息进行文献检索,在引文上下文自动识别上采用指定规则。

本文通过对相关研究进行全面的调研和总结,认为目前引文上下文识别研究存在 4 点不足: ①没有注意到一个引文句中包含多个引文标记的情况,这些引文标记是否有着独立的上下文以及它们之间的关系; ②将引文上下文定义为连续句子集合,且在确定候选上下文时将范围局限在引文标记周围的句子,忽略了引文上下文的复杂性; ③大部分研究者并未考虑非常完整的文献结构(如章节、段落等信息),而这些结构对引文上下文的识别也起着非常重要的作用; ④相关研究也缺乏对引文上下文特征的细致分析。针对上述不足,本文系统梳理了引文上下文的各种表现形式(引文上下文和引文标记在文章中的分布规律、跨句的引文上下文分析和多引用引文句的统计分析等),在引文上下文识别已有特征基础上增加了 4 类新特征(是否在同一段落、是否包含其他引文标记、前面句子的类型、后面句子的类型),并且在自构建的引文上下文语料基础上,运用 SVM 和 CRF 两种方法进行了引文上下文自动识别实验,取得了很好的效果。基于以上研究现状和研究思路的分析,本文提出的研究路线如图 1 所示:

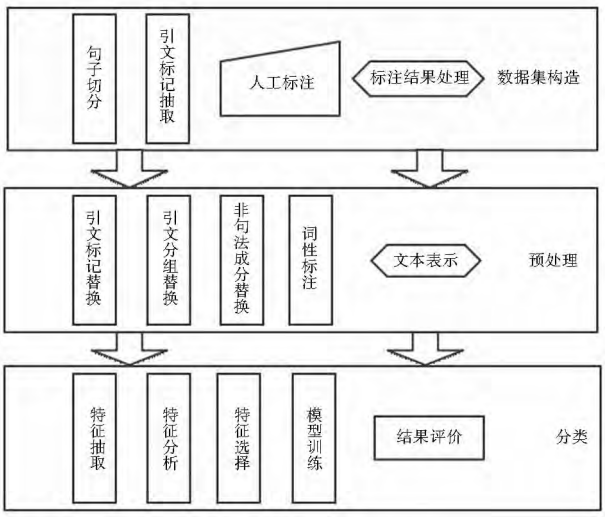

图 1 研究路线

3 引文上下文自动识别理论与方法

3.1 引文内容分析

引文分析就是利用各种数学模型和统计指标,对文献之间的引用和被引用现象进行统计分析,通过这些数据揭示文献引用的规律和特征,从而对文献进行评价和趋势分析的一种计量学方法。然而,传统的引文分析将所有引用关系同等对待,只能告诉读者哪两篇文章之间具有引用关系,不能说明被引成果对于施引文献的具体贡献以及重要性,这种对引用关系的简化处理无法展示科研文献网络的真实情景[3]。

引文内容分析通过对引文句及其上下文进行语法和语义分析,深入挖掘施引文献和被引文献之间的语义关联,揭示引用行为的本质,从而弥补了传统引文分析的不足。目前,引文内容分析主要包含以下几个方面的研究内容:

3.1.1 引文内容识别 引文内容识别是指在正文中寻找与引文相关的文本。只有得到了完整的引文内容,才能进行有效的引文内容分析。针对目标文献,首先需要查找该目标文献在整篇文章中所有的引用标记,然后在每个引用标记周围寻找与该引文相关的句子。引文内容识别是本文的主要研究内容,即引文上下文的自动探测,也是引文内容分析的基础。

3.1.2 引文功能分析 引文功能是指作者引用某篇参考文献的目的,也即被引文献在施引文献中起到的作用。例如作者将被引文献作为基础研究,在此基础之上进行算法的改进,或者将被引文献与其他文献进行对比以显示出不同,或者描述被引文献的局限性,还有就是作为相关研究内容仅仅列出被引文献的研究成果等。引文功能直观体现了参考文献在施引文献中的作用,因此在引文内容分析中处于核心地位。

3.1.3 引文情感分析 引用是研究者的主观行为,因此对被引工作的描述中也会包含一定的主观观点,这些描述可能是正面的、负面的和中立的,许多学者认为对不同情感的引文应当进行区分对待: 有些文献被引频次很高,但是大部分的施引文献都是对该文献进行批判的话,那么即使其有很高的被引,对该文献的评价也应给予一定的折扣。引文情感识别可以帮助研究者更准确地对文献进行评价,比如负面情感的文章在进行引文评价时可以不作加分或者作为减分项。对一篇文献的所有被引引文的上下文文本进行情感分析,可以用于改进学术文献检索,从而为用户提供更好的选择。引文情感分析直接体现了作者对于被引文献正面或负面的情感态度,也是学者在引文内容分析时较为关心的维度。

3.1.4 引文重要性分析 不同被引文献对于施引文献的重要程度不同,在进行引文分析的时候也要区别

对待,如果一篇文章的某些被引在施引文献中的重要程度非常低的话,可以适当降低权值。例如,有些引文仅用来标识概念、工具、方法等,而且有的时候这些概念、方法并不是被引文献的研究成果,也仅仅是提及或者使用,那么这个引文对于该文献的重要性就是比较低的。引文重要性反映一篇参考文献对于其施引文献智力支持程度的大小,能够帮助读者了解哪些被引成果在作者的研究中贡献了重要作用,因而是引文内容分析的重要方面。

引文内容分析使得更加准确科学地进行学术评价和研究热点分析成为可能,也带来很多基于引文内容分析的应用研究,例如学术语义搜索[21]、引文推荐[22-23]、自动摘要[24]等,因此受到学术界的广泛关注,越来越多的学者致力于探索更加高效的引文上下文自动识别方法。

3.2 引文上下文识别方法

随着自然语言处理和机器学习技术的发展,多种机器学习方法被用于引文上下文自动识别研究中。引文上下文识别既可以被看作短文本分类问题,也可以被看作序列标注问题,基于这两种思想,已有诸多学者运用各种方法对该问题进行了探索,笔者总结了具有代表性的引文上下文识别方法和特征见表1。

表1 引文上下文识别常用方法与特征

代表人物	方法	特征
K. Sugiyama 等	maximum entropy; support vector machine (SVM)	unigram(一元词袋特征); bigram(二元词袋特征); proper nouns(专有名词); previous and next sentence(引文标记的前后句子); position(位置特征); orthographic(字形特征)
V. Qazvinian 和 D. Radev	support vector machine (SVM)	citation features(引文特征); discourse-based features(基于语篇特征); sentence-level features(句子层面特征)
A. Abu-Jbara 和 D. Radev	support vector machine (SVM)	similarity to the target paper(与目标文献的相似度); headlines(标题); relative position(相对位置); first person pronouns(第一人称代词); tense of the first verb(第一个动词时态); determiners(限定词)
A. Abu-Jbara 和 D. Radev	word classification; sequence labeling; segment classification	distance(距离); position(位置); segment(片段); part of speech tag(词性标签); dependency distance(依存距离); dependency relations(依存关系); common ancestor node(共同祖先节点); syntactic distance(语义距离)
M. A. Angrosh 等	conditional random fields (CRF)	citation features(引文特征); section features(章节特征); term features(词汇特征)
P. Sondhi 和 C. X. Zhai	hidden Markov model (HMM)	

3.2.1 基于文本分类的上下文识别 引文上下文的识别可以通过文本分类的思想来进行。常用的文本分类[25]方法包括朴素贝叶斯方法(Naïve Bayes classifier)、K近邻(K-nearest neighbor classifier)、决策树、支持向量机(SVM)等。

本文使用SVM对引文上下文进行分类实验。针对隐式引文上下文识别采用二值分类来解决,即对候选上下文语句的识别结果只有两种:引文上下文(类别为1)、非上下文(类别为0)。具体识别步骤为:

步骤1:数据预处理。对语料数据进行预处理,主要包括文本格式数据的获取、句子切分、分词、词法分析等。另外,还包括针对文献数据特有的引文标记进行处理的过程。本文在预处理中将特征抽取中需要用到的所有信息都进行分析并以特定的格式存储。

步骤2:特征抽取和特征选择。特征构造是分类问题中比较重要的部分,特征质量的好坏直接影响分类器的效果。本文在系统分析前人成果的基础上,结合实例分析定义了若干组新特征。随后编写程序抽取所有特征并进行特征选择。特征选择可以对特征集中的大量特征进行分析并保留其中最为有效的特征子集。如在本文中,N-Gram会产生大量的特征,特征维度过高会影响分类器的速度和准确度,因此可以通过特征选择提高分类的效果。

步骤3:SVM模型的训练和评价。使用SVM对前面两个步骤中抽取的特征进行分类。首先要对参数g和c进行迭代实验,选择出效果最好的g、c参数,然后使用这两个参数对数据进行训练和预测,并对预测的结果进行评价。

3.2.2 基于序列标注的上下文识别 序列标注可以看作是分类的推广。常用的序列标注模型有隐马尔科夫模型(hidden Markov model,HMM)和条件随机场(CRF)。关于这两种模型的介绍,请参见文献[26]。

本文选择CRF进行引文上下文的识别,原因是CRF可以找到整个序列上的全局最优,且常被用于文本分词、词性标注、组块识别、命名实体识别等序列标注类任务中。使用CRF进行上下文识别的过程如下:①预处理。该过程和使用SVM进行预处理的过程类似。②特征抽取。CRF是序列标注问题,所以在特征抽取时会得到一个序列,即对于每一个引文标记,它的所有候选上下文依照句子顺序得到的标注结果构成了一个标记序列。③序列标注。使用序列标注工具上一

步抽取的特征文件进行处理，得到每一个序列标注的标注结果，然后对标注结果进行查全、查准评价。

本文之所以选择 SVM 和 CRF 进行实验，一方面是因为这两种方法已被证明在短文本分类和序列标注中十分有效，另一方面是因为本文的创新点之一为引文上下文识别特征的改进，笔者期望将提出的引文上下文识别特征融入最优的机器学习模型中，实现引文上下文识别效果的提升。

3.3 引文上下文识别特征

笔者结合已有相关研究成果[12-13, 16, 27]并对引文句及其周围句子的内容和结构进行细致分析，总结了 5 类引文上下文识别的有效特征，如表 2 所示：

表 2 引文上下文识别的特征

特征类别	子特征	解释
指代特征	◇是否包含 Work Nouns	句子中是否包含了 Work Nouns 词组
	◇是否包含引文句中目标引文标记前面相邻的名词短语	句子中是否包含了目标引文标记的某个指代词语
	◇是否包含作者名字	句子中是否包含目标引文中作者的名字
	◇是否包含第三人称代词	句子中是否包含指定的第三人称指代词语
	◇Lexical hooks:(词表)	句子中是否包含 Lexical hooks 词汇
位置特征	◇相对位置	与目标引文句的距离
	◇段落标记特征	是否和目标引文在同一个段落
	◇Section 特征:	
	特征 1	句子是否是章节的第一句
	特征 2	句子是否是章节的最后一句
	特征 3	当前句的前一句是否为章节的第一句
	◇前后句子的类型:	
	特征 1	当前候选上下文句前面一句是否为非目标引文句，如果前面一句是对其他文章的引用，则该句不大可能是目标引文的上下文
	特征 2	当前候选上下文句后面一句是否为目标引文句，如果该句后面相连的句子引用了其他文献，则该句不大可能是目标引文的上下文
	◇该句包含引文标记但不包含目标引文	当前句中是否只包含其他引文，若包含了其他引文的引用，则该句属于目标引文上下文的可能性会降低
	◇在文章中的区域	句子在文章中的区域
句子结构特征	◇连接副词	候选引文上下文句是否起始于指定的连接副词，如 However、Therefore 等
内容特征	◇1-3grams	
	特征 1	1-gram
	特征 2	2-grams
	特征 3	3-grams
	◇1-3grams 相似度	
	特征 1	1-gram 相似度
	特征 2	2-grams 相似度
	特征 3	3-grams 相似度
类型特征	◇引文句中的引文个数	引文句中的引文标记个数

为了验证上文提出的特征方案的科学性，本文基于自构建的数据集(对原始文本进行处理而得到的数据，具体构建过程见 4.1 和 4.2 节)对引文标记包含的上下文个数、引文上下文和目标引文的距离、引文上下文在文章整体位置的分布、跨句的上下文类型、多引用引文句中引文标记的上下文分布、引用独立性等能表征引文上下文特征的信息进行了统计分析，很好地验证了特征方案的可行性。

4 实验与结果分析

4.1 数据集

4.1.1 数据来源 本文所选用的原始数据集是计算语言协会(Association for Computational Linguistics, ACL)的网络语料库选集[28]，该选集包含了 34 000 多篇计算语言学和自然语言处理相关的期刊和会议论文集中的英文文献。本文使用的数据集是 U. Schäfer 等[29]通过 OCR 等技术对 ACL 数据集进行识别而得到的结构完整的 XML 文档数据。该文档包含完整的章节、段落和参考文献等信息。本实验随机抽取了其中 130 篇文档。

4.1.2 数据标注 针对本研究，笔者开发了专门的引文标注系统(用于标注引文句及其上下文范围，得到训练数据集)进行引文上下文的标注(见图 2)。标注以句子为单位，句子被分成 3 类：引文句(包含引文标记的句子)、普通句(没有引文标记的句子)、上下文句(需要标注的类型)。另外引文句也会被标注为上下文句。

笔者邀请武汉大学信息管理学院 13 名从事自然语言处理和信息检索研究的研究生对文献进行标注，具体工作如下：

(1)确保引文标记抽取的正确性：如果有错误的引文标记，则直接将该句子的类型改为普通句。

(2)确保分句的正确性：如果分句有错误则进行标记(T 表示向上合并、B 表示向下合并)，后期数据处理时会进行合并。

(3)引文上下文标注：标注者需要识别每一个引文标记所在章节内的所有句子，判断该句是否属于目标引文标记的上下文。如果是，则将该句标记为上下文句，并在对应的文本框位置填写该引文标记对应的序号，多个序号用指定字符分隔。本文第一作者对所有文章进行了标注，以保证标注结果的可靠性。

4.1.3 一致性检验 本文首先对标注者标注结果的一致性进行评估，结果一致性的高低体现了本标注系

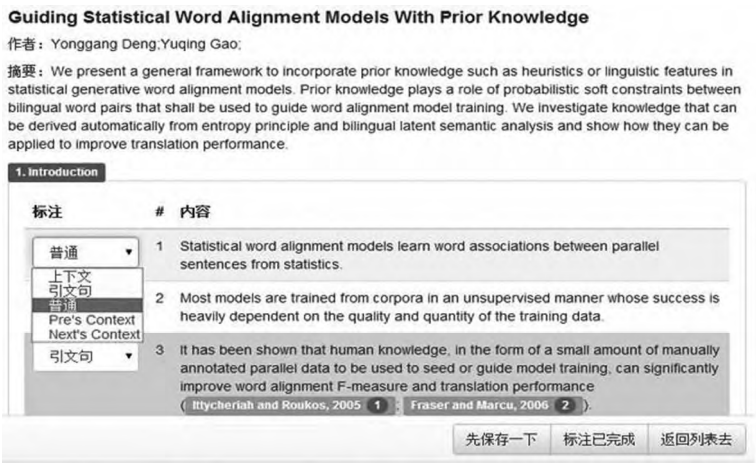

图2 引文上下文标注系统

统的可操作性和训练集的可靠性。本文重点对上下文范围的标注结果进行一致性评估和分析。Kappa 系数是一个被广泛使用的一致性评价机制,其计算公式如下:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

其中 P(A) 表示标注结果一致性的实际观测值,P(E) 表示标注结果一致性的期望值。本文选取第一作者和其中一位标注者共同标注的 20 篇文献(共包括 3 700个句子)做交叉检验,标注结果显示 Kappa 一致性为 $K = 0.937$,说明该标注结果达到了较为可靠的一致性水平。

得到标注数据后,需要对标注结果的 XML 文档进行处理,主要是对标注结果进行校正和规范化,如句子合并、删除,人工处理没有抽取的引文标记和抽取错误的引文标记。另外,遍历所有引文标记(ref 标签),并给该标签增加 context 属性,context 属性的内容为该引文标记所有上下文的句子编号。最后得到 130 篇包含上下文内容标记的 XML 格式语料文献。

本文的特征还要用到很多词表,如第三人称代词、连接副词、WorkNouns、Lexicalhook 停用词表等。笔者随机从 130 篇语料集中挑选 30 篇作为发展集(develop set,用于统计连词、人称代词、指示词、对象词等词表的数据集),然后统计这些数据。收集完的词表用作特征抽取。

4.2 数据预处理

为了有效实现引文上下文识别,数据预处理过程十分必要,预处理是为特征抽取做准备。特征抽取需要用到的一系列词汇信息、统计信息等都可以在预处理过程中完成。首先本文以句子为单位对文献进行研究,因此需要对文章内容进行分句;由于现有的词性标注和句法分析工具仅在一些标准的语料库中进行训练,无法对文献中的引文标记进行处理,而文献中包含大量引文标记,这些标记有的占一定的句法成分,因此笔者认为有必要对引文标记进行处理,以保证句子在词法和句法上的规范性,保证词法和句法工具能够得到更准确的结果;本文中还需要用到词性信息,因此需要对句子进行词性标注。

4.2.1 句子切分 上面提到,本研究是以句子为基本单位开展的,需要对段落进行分句。本文使用 OpenNLP Tools[8] 进行句子切分,该工具支持多种语言。句子切分在语料标注之前完成。英文一般是通过句点进行句子分隔,但由于句点在英文中还有很多含义,如放在缩写词后面等,故可能会产生错误的句子切分,因此本文在标注时也通过人工进行错误分句的合并和无效句子的删除。

4.2.2 引文标记处理 考虑到引文标记的复杂性,笔者对引文标记的处理分以下 3 步进行:

(1) 引文标记的替换。引文标记一般通过内嵌的形式或括号形式加到句子中,引文标记使用特定的格式在语法上是不标注的,因此需要对引文标记进行替换。本文数据集中的引文标记属于比较通用的格式,主要有两种形式:①“作者列表(年份)”;②“(作者列表,年份)”。多个引用标记会用分号隔开。作者是首字母大写的词语,作者列表之间通过“&”符号或“and”分隔或“et al.”进行省略。年份是 4 位数字。引文标记出现的位置主要有 3 种情况:①在某个名词或名词短语后,代表该名词或名词短语的参考文献;②在句子结尾,代表该句的参考文献;③作为句子的句法成分出现,这种情况一般是对该参考文献直接进行介绍。基于上面对引文标记的分析,笔者首先通过正则表达式匹配出所有的引文标记,并直接将其替换为 REF,通过设计良好的数据结构保留了 REF 对应的引文内容,不会丢失任何信息;然后删除引文标记相邻的括号并记录到引文标记对应的数据结构中。

(2) 相连引文标记整体替换。很多引文句包含了不止一个引文标记,这些并列出现的引文标记大多是作为整体来进行引用的,故需要进行引文分组替换。笔者将多个连续的 REF 合并为一个 GREF 并记录了该 GREF 中保留的所有信息。

(3) 非句法成分替换。引文标记不一定属于句法的一部分,如果不属于句法成分,则会对词法和句法分

析产生干扰，故需要将不属于句法成分的 REF 或 GREF 移除。这些不属于句法成分的标记一般是对它前面的一些专有名词的注解，表示该词的来源，因此，移除的标记会被附着到其指代词语上。本文采用了基于规则的方法判断标记类型，如果属于句法成分则保留，否则就找到标记前面的词语代表该标记。具体的规则描述如下：①如果引文标记符合“作者(年份)”模式，则该引文标记肯定是句法成分，保留该标记；②如果引文标记在第一个位置，则肯定是句法成分，保留该标记；③如果引文标记前面的词语为指定介词(如 in、by、with、like 等)，但不一定是介词，则认定该引文标记是句法成分，一般会充当宾语等成分，保留该标记；④其他情况下，移除该标记并将该标记的信息放置到其前一个词语中。

用一个实例来描述以上对引文标记的处理过程，如图 3 所示：

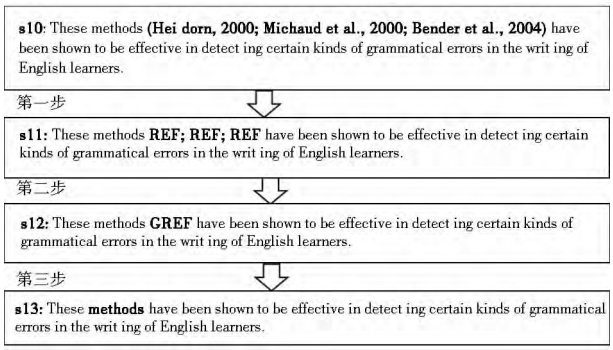

图 3 引文标记处理实例

经过上面 3 步处理，可以发现，s13 是一个比较标准的句子，在进行词法和句法分析中会得到更好的效果。

4.2.3 词性标注 引文标记处理完成后，就可以使用词性标注工具处理，然后将标注结果存储下来供特征抽取和分类使用。

本文语料均为英文，虽然英文有天然的空格分词，但由于文本中包含大量的标点符号无法通过空格进行拆分，因此笔者采用正则表达式进行分词。

分词完毕后得到一个词序列，然后使用 Stanford Parser(斯坦福语法解析工具)[31] 的词性标注工具进行词性标注，本实验需要用到的词性主要有介词(IN)、代词(PR)、名词(NN)、连词(CC)等，其中介词用来进行非句法成分替换中词汇的搜集，其他几类词语都用在了特征抽取中。

4.3 特征抽取

4.3.1 特征分析 预处理完成后，通过程序进行特征抽取，实验选择的特征见表 2。然后笔者对选取的特征做重要分析，实验使用 Weka[32] 的特征选择工具进行特征分析，使用信息增益来对特征进行排序。信息增益是对增加某个特征而使类别信息的不确定性减少的程度的度量。表 3 是对本实验选取的 19 个特征按照信息增益进行排序的结果。

表 3 实验特征及其信息增益

编号	特征	类型	信息增益
1	与目标引文的距离	位置	0.242 230 7
2	是否在同一段落	位置	0.197 598 0
3	是否包含其他引文标记	位置	0.115 275 7
4	前面的句子类型为非目标引文句	位置	0.100 523 7
5	1 - gram 相似度	内容	0.081 457 2
6	2 - gram 相似度	内容	0.050 127 2
7	后面句子为非目标引文句	位置	0.033 497 8
8	3 - gram 相似度	内容	0.026 148 0
9	前一个句子是否章节第一句	位置	0.005 968 2
10	是否包含引文指代的词语	指代	0.005 040 2
11	包含引文标记的个数	类型	0.004 249 1
12	是否包含第三人称代词	指代	0.004 035 1
13	是否包含指定连词	结构	0.001 806 8
14	是否包含作者名称	指代	0.001 107 0
15	是否是章节第一个句子	位置	0.001 006 1
16	是否是章节最后一个句子	位置	0.000 770 2
17	在文章中的区域	位置	0.000 445 6
18	是否包含 Work Nouns	指代	0.000 088 3
19	是否包含 Lexical Hooks	指代	0.000 056 1

从表 3 可以发现，位置特征的信息增益较高，说明上下文还是围绕在引文标记周围的(特征 1、2)，且有一定的连续性(特征 4、7)；其次是内容特征，说明引文句和上下文句在内容上具有一致性；然后是指代特征，说明了这 3 个特征在分类中的有效性(特征 10、12、14)。其中，黑体标明的特征(特征 2、3、4、7)是笔者提出的特征，它们的信息增益都较高，笔者将在后面实验中验证其有效性。

4.3.2 N - Gram 特征选择 N - Gram 特征抽取中会产生大量特征，必须对这些特征进行词频过滤(选择词频大于 5 的词条)得到部分子集，然后使用信息增益对该子集进行选择。

本实验首先从范围为 100 - 1 000 的特征数中选择了 10 组特征值，然后对每组特征值进行分类实验，选择结果最好的一组作为 N - Gram 最终特征。

4.4 引文上下文自动识别实验

4.4.1 基于 SVM 的分类实验 该实验使用 LibSVM[33] 进行参数选择和分类实验，使用 10 倍交叉

检验进行参数训练。笔者首先使用 LibSVM 的 grid 工具进行参数训练,默认将 g 的范围设为 [-12,2],c 的范围设为 [0,12],步长设为 0.2,通过观察训练结果,不断调整参数,得到最好的 SVM 参数后采用 svm-train 进行 10 倍交叉检验。实验选取的候选上下文为同一章节前后各 4 句,其依据在前面已做论证。然后编写程序进行特征抽取,共得到正例(类别标签为 1)3 578 个,负例(类别标签为 0)20 776 个。为保证正例和负例平衡,笔者随机抽取负例 3 480 个进行分类实验。本实验选取 3 组对照实验,分别如下:

(1)使用内容特征(SVM_N-Gram)用 SVM 训练。该组实验仅使用文本内容进行上下文识别,选择 1-3gram 这 3 个特征进行上下文分类。

(2)使用 15 类特征(SVM_F15)用 SVM 训练。该组实验共有包含文献调研中所获的 15 个特征参加分类实验,然后通过参数训练和分类得到结果。

(3)使用全部特征(SVM_F19)用 SVM 训练。该组实验包含表 2 的所有特征进行分类实验,通过对比上一组实验来确认笔者所添加特征的有效性。实验结果见表 4。每个类别的实验结果见表 5。

表 4 3 组实验分类结果

实验	P	R	F1
SVM_N-Gram	0.633 473	0.592 410	0.556 113
SVM_F15	0.783 863	0.782 000	0.782 013
SVM_F19	0.856 571	0.856 526	0.856 333

表 5 分类在各个类别上的结果

类别	总数	实验	正确数	P	R	F
1	3 578	SVM_N-Gram	1 123	0.436 117	0.714 377	0.313 862
		SVM_F15	2 931	0.767 478	0.819 173	0.792 483
		SVM_F19	3 015	0.869 879	0.842 650	0.856 048
0	3 480	SVM_N-Gram	3 031	0.552 497	0.870 977	0.676 110
		SVM_F15	2 592	0.800 247	0.744 828	0.771 543
		SVM_F19	3 029	0.843 263	0.870 402	0.856 618

表 4 中,基准实验仅选择内容特征进行分类,在特征数为 800 时取得最好结果 0.556 113,但该结果并不理想。这说明单使用内容特征无法对上下文进行很好的分类,虽然隐式上下文和引文句在内容上关联,但上下文相比引文句描述了更多引文相关的内容。实验 F15 在 Baseline 的基础上取得很大提升,其 F 值达到 0.78,表明选取的这些特征能够使实验取得不错的效果。最后,在增加笔者提出 4 个特征的 F19 实验中,F 值达到0.856,大大提升了分类效果,证明笔者提出的 4 个特征对分类是十分有效的。

表 5 列出了 3 组分类实验在各个类别的实验结果。从中可知,在各个类别上,准确度、召回率和 F 值都是一次递增的,且 F19 相对 F15 在两个类别上 F 值均取得了很大的提升。

4.4.2 基于 CRF 的引文上下文序列标注实验 该实验使用了 CRF++[34] 进行序列标注,使用 5 倍交叉检验验证实验自动识别的准确率。笔者首先进行了特征抽取,将同一引文的候选引文上下文相关特征作为一个序列抽取出来,共得到 3 719 组序列数据,然后将这 3 719组序列分为 5 等份进行分组实验,将其中 1 组作为测试集,另外 4 组作为训练集。本实验选取了 2 组对照实验,分别如下:

(1)使用 15 类特征(CRF_15)用 CRF 训练。该组实验共有包含文献调研中所获的 15 个特征参加分类实验,然后进行序列标注实验。

(2)使用全部特征(SVM_F19)用 SVM 训练。该组实验包含表 2 的所有特征进行分类实验,通过对比上一组实验来确认笔者所添加特征的有效性。实验结果见表 6。每个类别的实验结果见表 7。

表 6 3 组实验分类结果

实验	P	R	F1
CRF_F15	0.726	0.697	0.709
CRF_F19	0.822	0.799	0.808

表 7 分类在各个类别上的结果

类别	总数	实验	正确数	P	R	F1
1	3 578	CRF_F15	1 763	0.597	0.494	0.540
		CRF_F19	2 387	0.740	0.669	0.700
0	11 983	CRF_F15	10 792	0.856	0.900	0.877
		CRF_F19	11 137	0.903	0.930	0.916

表 6 中,实验 CRF_F15 作为 Baseline 取得了 0.709 的 F1 值,而 CRF_F19 则达到 0.808,超出 Baseline 近 10%,这说明笔者添加的 4 个特征对序列标注的效果影响是非常大的。

表 7 中列出了每个类别上的实验结果。可以发现,类别 0 的 F1 值相比类别 2 高出很多,这可能是由于类别 0 的实例过多导致的。

本节笔者使用两种方法进行了引文上下文识别,第一种方法使用了文本分类思想,选择 SVM 进行上下文识别;第二种方法使用了序列标注思想,将候选上下文作为一个序列进行标注,选择 CRF 进行上下文类别的标注。实验结果发现,基于 SVM 的文本分类效果要好于 CRF。当然,一方面是由于两者使用的训练集有差

别,另一方面,在大量测试实验的基础上,笔者认为将序列标注思想用到引文上下文自动识别还有待进一步探讨。

5 总结与展望

本文首先由引文内容分析引入引文上下文,然后通过文献调研提出了针对引文上下文自动识别的任务,接着介绍了相关分类方法,之后总结了引文上下文识别的 5 类特征。最后,使用文本分类模型 SVM 和序列标注模型 CRF 开展了引文上下文自动识别实验,取得了很不错的效果,同时也验证了笔者添加分类特征的有效性。但仍有一些有待改进的地方:

(1) 针对每篇文献的标注结果做有效性分析。笔者在实验中发现有相当部分标注结果存在问题,若能对这些数据进行修正,将会进一步提升实验的效果。

(2) 本文在预处理时做了引文标记和参考文献配对工作,但在后续工作中并未对相关内容做深入研究。

在下一步工作中,笔者将对本研究的不足加以改进,并将引文上下文自动识别的结果应用到基于引文上下文的学术文献引文推荐和自动摘要中。

参考文献:

[1] 刘洋,崔雷. 引文上下文在文献内容分析中的信息价值研究[J]. 图书情报工作,2014,58(6):101 - 104.

[2] ABU-JBARA A, EZRA J, RADEV D R. Purpose and polarity of citation: towards NLP-based bibliometrics [C] //Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. Atlanta: Association for Computational Linguistics, 2013: 596 - 606.

[3] 陆伟,孟睿,刘兴帮. 面向引用关系的引文内容标注框架研究[J]. 中国图书馆学报,2014(6):93 - 104.

[4] COLLINS H M. The TEA set: tacit knowledge and scientific networks [J]. Social studies of science, 1974, 4(2): 165 - 185.

[5] CANO V. Citation behavior: classification, utility, and location [J]. Journal of the American Society for Information Science, 1989, 40(4): 284 - 290.

[6] CHUBIN D E, MOITRA S D. Content analysis of references: adjunct or alternative to citation counting? [J]. Social studies of science, 1975, 5(4):423 - 441.

[7] NANBA H, OKUMURA M. Towards Multi-paper summarization using reference information [C] // Proceedings of The 1999 International Joint Conference on Artificial Intelligence. Stockholm: AAAI, 1999: 926 - 931.

[8] ABU - JBARA A, RADEV D. Coherent citation-based summarization of scientific papers [C] //Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies-volume 1. Portland: Association for Computational Linguistics, 2011: 500 - 509.

[9] ATHAR A. Sentiment analysis of citations using sentence structure -based features [C] //Proceedings of the ACL 2011 student session. Portland: Association for Computational Linguistics, 2011: 81 - 87.

[10] ANGROSH M A, CRANEFIELD S, STANGER N. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries [C] //Proceedings of the 10th annual joint conference on digital libraries. Gold Coast: ACM, 2010: 293 - 302.

[11] QAZVINIAN V, RADEV D R. Identifying non-explicit citing sentences for citation-based summarization [C] //Proceedings of the 48th annual meeting of the association for computational linguistics. Uppsala: Association for Computational Linguistics, 2010: 555 - 564.

[12] KAN M Y. Identifying citing sentences in research papers using supervised learning [C] //2010 International conference on information retrieval & knowledge management (CAMP). Toronto: IEEE, 2010: 67 - 72.

[13] ABU-JBARA A, RADEV D. Reference scope identification in citing sentences [C] //Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies. Montréal: Association for Computational Linguistics, 2012: 80 - 90.

[14] ANGROSH M A, CRANEFIELD S, STANGER N. Conditional random field based sentence context identification: enhancing citation services for the research community [C] //Proceedings of the first Australasian Web Conference-Volume 144. Adelaide: Australian Computer Society, 2013: 59 - 68.

[15] SONDHI P, ZHAI C X. A constrained hidden Markov Model Approach for Non-Explicit Citation Context extraction [C] // Proceedings of the 2014 Society for Industrial and Applied Mathematics International conference on data mining. Pennsylvania: Society for Industrial and Applied Mathematics, 2014: 361 - 369.

[16] ATHAR A. Sentiment analysis of scientific citations [EB/OL]. [2016 - 05 - 10]. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-856.pdf.

[17] 刘盛博,丁堃. 基于引用内容的引文评价分析[C] //第九届中国科技政策与管理学术年会论文集. 济南:山东省科技发展战略研究所,2013.

[18] 许德山. 科技论文引用中的观点倾向分析[D]. 北京:中国科学院文献情报中心,2012.

[19] 孙枫军. 引文上下文中的概念抽取[D]. 北京:中国科学信息技术研究所,2012.

[20] 张金松. 基于引文上下文分析的文献检索技术研究[D]. 大连:大连海事大学,2013.

[21] SCHAFER U, SPURK C. TAKE scientist's workbench: semantic search and citation-based visual navigation in scholar papers[C]//IEEE International conference on semantic computing. Pittsburgh: IEEE,2010:317-324.

[22] TANG X, WAN X, ZHANG X. Cross-language context-aware citation recommendation in scientific articles[C]// Proceedings of the 37th International ACM SIGIR conference on research & development in information retrieval. Gold Coast: ACM, 2014:817-826.

[23] LIVNE A, GOKULADAS V, TEEVAN J, et al. CiteSight: supporting contextual citation recommendation using differential search[C]// Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. Gold Coast: ACM, 2014:807-816.

[24] COHAN A, GOHARIAN N. Scientific article summarization using citation-context and article's discourse structure[C]// Conference on empirical methods in natural language processing. Lisbon: Association for Computational Linguistics,2015.

[25] 杨杰明. 文本分类中文本表示模型和特征选择算法研究[D]. 长春:吉林大学,2013.

[26] 鉴萍,宗成庆. 基于序列标注模型的分层式依存句法分析方法[J]. 中文信息学报,2010,24(6):14-22.

[27] ATHAR A, TEUFEL S. Detection of implicit citations for sentiment detection[C]//Proceedings of the workshop on detecting structure in scholarly discourse. Jeju Island: Association for Computational Linguistics, 2012: 18-26.

[28] RADEV D R, MUTHUKRISHNAN P, QAZVINIAN V. The ACL anthology network corpus[C]//Proceedings of the 2009 workshop on text and citation analysis for scholarly digital libraries. Stroudsburg: Association for Computational Linguistics, 2009: 54-61.

[29] SCHÄFER U, WEITZ B. Combining OCR outputs for logical document structure markup: technical background to the ACL 2012 contributed task[C]//Proceedings of the ACL-2012 special workshop on rediscovering 50 years of discoveries. Jeju Island: Association for Computational Linguistics, 2012: 104-109.

[30] [EB/OL]. [2016-05-10]. http://opennlp.apache.org/ to download OpenNLP.

[31] [EB/OL]. [2016-05-10]. http://nlp.stanford.edu/software/lex-parser.shtml.

[32] [EB/OL]. [2016-05-10]. http://www.cs.waikato.ac.nz/ml/weka/.

[33] [EB/OL]. [2016-05-10]. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[34] [EB/OL]. [2016-05-10]. http://wing.comp.nus.edu.sg/~forecite/services/parscit-100401/crfpp/CRF++-0.51/doc/.

作者贡献说明:

雷声伟:参与研究思路修改讨论和研究框架整理,参与实验,撰写和修改论文;

陈海华:参与论文研究框架整理,进行文献调研,修改论文;

黄永:负责数据处理,参与实验和论文修改;

陆伟:提出研究思路和研究整体框架,参与论文修改。

Research on Automatic Recognition of Academic Citation Context

Lei Shengwei[1] Chen Haihua[1] Huang Yong[1] Lu Wei[1,2]

[1]School of Information Management, Wuhan University, Wuhan 430072

[2]Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072

Abstract: **[Purpose/significance]** Citation content analysis can help to reveal the deep semantic influence of literature citation relations, and citation context identification as a basis for content analysis is particularly important. **[Method/process]** This paper reviews the latest development of researches of citation context and summarizes the deficiencies in citation context identification. Based on which five categories of citation context identification features are proposed. Besides, this paper also conducts an automatic identification experiment by utilizing text classification and sequence labeling. **[Result/conclusion]** A significant improvement over baseline method shows the effectiveness of our features. Besides, the text classification based SVM method performs better than the sequence labeling based CRF method.

Keywords: citation context citation analysis support vector machine condition random field no-explicit context