文本蕴含关系识别与知识获取研究进展及展望

郭茂盛 张 宇 刘 挺

(哈尔滨工业大学计算机科学与技术学院社会计算与信息检索研究中心 哈尔滨 150001)

摘 要 文本蕴含关系是广泛分布于自然语言文本中的单向推理关系,文本蕴含相关研究是自然语言处理领域的一项基础性研究,它可以辅助其他自然语言处理任务的进行,并且具有丰富的应用场景.文中首先界定了文本蕴含研究的范畴.作为一种二元关系,文本蕴含有3个基本研究任务——关系识别、知识获取和蕴含对生成.其中,关系识别有两个核心问题——语义表示与推理机制;知识获取也有两个核心问题——知识表示与知识来源;蕴含对生成研究进展比较缓慢,文中细致地分析了其内因和外因.文中围绕语义表示与推理机制这两个核心问题梳理了关系识别的研究进展,围绕知识表示与知识来源梳理了知识获取的研究进展,并指出了各类方法的可取之处与不足之处.文本蕴含研究的进展离不开相关国际评测,文中也对这些国际评测和数据集进行了归纳总结.大数据时代的到来和深度学习理论的不断发展,为文本蕴含相关研究提供了丰富的知识来源和有力的研究工具,同时也带来了许多崭新的研究课题.文中立足当前研究形势,展望了未来研究方向,并从理论上探讨了其可行性.

关键词 文本蕴含;知识获取;自然语言理解;自然语言处理;人工智能

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2017.00889

Research Advances and Prospect of Recognizing Textual Entailment and Knowledge Acquisition

GUO Mao-Sheng ZHANG Yu LIU Ting

(*Research Center for Social Computing and Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract Textual entailment, as a directional semantic reasoning relation, is widely distributed in natural language texts. Research on textual entailment is a fundamental study in the field of natural language processing. With various applications, it is helpful to other natural language processing tasks. This paper clarifies the scope of textual entailment at first. As a binary relationship, textual entailment has three basic research tasks, that is, recognizing textual entailment, knowledge acquisition and generating entailment pairs. There are two key problems in recognizing textual entailment, that is, semantic representation and reasoning mechanism. There are also two key problems in knowledge acquisition, that is, knowledge representation and knowledge source. This paper makes a detailed analysis on the internal and external factors leading to the slow process of research on generating entailment pairs. This paper focuses on these key problems while expounding methods of recognizing textual entailment and knowledge acquisition. This paper points out the pros and cons of each method then. The development of research on textual entailment is inseparable with international evaluation exercises. This paper summarizes the datasets from these evaluation exercises. The arrival of the big data era and the development

收稿日期:2016-04-21;在线出版日期:2016-10-11.本课题得到国家自然科学基金(61472105,61472107)、国家"八六三"高技术研究发展计划项目基金(2015AA015407)资助.**郭茂盛**,男,1991年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为文本蕴含.E-mail:msguo@ir.hit.edu.cn.**张 宇**,男,1972年生,博士,教授,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为个性化信息检索、问答.**刘 挺**(通信作者),男,1972年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为社会计算、信息检索、自然语言处理.E-mail:tliu@ir.hit.edu.cn.

of deep learning theory bring a new rich source of knowledge and powerful tools, as well as novel research topics. The future research directions are pointed out and their feasibility is also discussed under the current research situation.

Keywords textual entailment; knowledge acquisition; natural language understanding; natural language processing; artificial intelligence

1 引 言

1.1 文本蕴含的研究背景

随着自然语言处理(Natural Language Processing,NLP)领域研究的不断深入,如何让机器能够真正地理解自然语言,而不是仅仅简单地处理语句的表层信息,渐渐成为了许多学者面临的问题.实现对文本深层次理解,是自然语言处理研究最主要也是最重要的目的之一.如果将其比作是自然语言处理研究领域的一顶皇冠的话,那么基于自然语言的语义推理无疑是这顶皇冠上最璀璨的一颗明珠.因为在获取了文本的语义后,一旦获得了它们之间的推理关系,这些文本便不再互相孤立,而是彼此联系起来,构成一张语义推理网络,从而促使机器能够真正理解并应用文本的语义信息.

文本间的推理关系,又称为文本蕴含关系[1](Textual Entailment,下一节将给出详细定义),作为一种基本的文本间语义联系,广泛存在于自然语言文本中.很多自然语言处理任务或多或少地都需要面对包含蕴含关系的文本,如果有一种技术能够识别其中的蕴含关系,那这种技术就能够为这些任务提供助力.因此,文本蕴含相关研究是自然语言处理领域的一项基础性工作.

例如,在问答系统中,若提问"谁是网易公司的创始人?",而语料库中恰恰有诸如"丁磊于 1997 年 5 月创立了网易公司."这样的句子,如果问答系统能够由此推理得出"丁磊是网易公司的创始人"的话,就可以直接对这样的问题进行作答.事实上,问题与候选答案、候选答案与支持文档之间一般都存在推理蕴含关系.有些问答系统[2-5]便利用文本蕴含技术来生成候选答案,或对用其他方法生成的候选答案进行筛选排序.实验[5]表明,应用文本蕴含技术能够把回答正确率提高 20%左右.

在关系抽取领域中,Romano 等人[6]使用文本蕴含技术扩展了抽取所用的模板,极大地丰富了目标关系的表现形式,从而提升了抽取的召回率.

在多文档文摘任务中,候选文摘中句子间的蕴含关系一定程度上指示了它们之间的语义包含关系,因此可以使用文本蕴含技术来辅助精简文本[7].

在机器翻译评价领域,文本蕴含技术也有一席之地.在理想情况下,正确的机器译文应当和人工标注的标准答案具有相同的语义,因而双方彼此可由对方文本推理得出,所以可以利用机器译文和标准译文的互相蕴含程度来对机器翻译系统的性能进行评估.Padó 等人[8]据此建立了基于文本蕴含技术的机器翻译自动评价系统.

类似地,在学生作业评分任务[9]中,学生的作答与标准答案之间的蕴含关系也可以指示学生答案的完善程度.Nielsen 等人[10]据此利用文本蕴含技术建立了一套学生作业评分系统.

在句法分析结果评价领域中文本蕴含技术也有用武之地.由于错误的句法分析结果会导致蕴含关系判定失败,因此可以使用基于句法特征的文本蕴含识别系统对其进行反向评价[11].

同时,在人们的日常生活中,近年来出现了不少所谓的"个人智能助理",例如 Apple Inc. 的 Siri 语音助手,Microsoft 的小娜(Cortana)个人助理等.她们能够聆听并"理解"用户的一些简单命令,帮助用户处理一些日常生活的简单任务,从而提高了用户的工作效率,也增加了这些智能设备的可玩性.但是,当前的个人智能助理并不能很好地处理用户的复杂需求,也不能在回答用户问题时有效地举一反三.其技术瓶颈在于当前技术不能有效地理解用户的语义并进行推理,这与目前文本蕴含相关技术尚未达到成熟商用的水平有关.因此,研究文本蕴含相关技术是日常生活应用的迫切需要.

文本蕴含相关研究的终极目标就是提供一个一般意义上基于文本的推理引擎来支撑其他语义相关的自然语言处理任务以及日常应用.

1.2 文本蕴含的研究范畴

1.2.1 文本蕴含关系的定义

文本蕴含的概念由 Dagan 等人[1]于 2004 年首次提出,其定义如下.

定义 1.　文本蕴含定义为一对文本之间的有向推理关系，其中蕴含前件记作 T(Text)，蕴含后件记作 H(Hypothesis). 如果人们依据自己的常识认为 H 的语义能够由 T 的语义推理得出的话，那么称 T 蕴含 H，记作 T⇒H[1].

T1：丁磊 1997 年 5 月创立网易公司.
H1：丁磊是网易公司的创办人.
T2：丁磊 1997 年 5 月创立网易公司.
H2：丁磊不是网易公司的创办人.
T3：丁磊 1997 年 5 月创立网易公司.
H3：丁磊是个中国人.

举例来说，T1-H1 符合前述文本蕴含的定义，它们的关系称为阳性蕴含关系(Positive Textual Entailment). 在不引起歧义的情况下，可以将阳性蕴含关系简称为蕴含关系；对于 T2-H2，人们获知 T2 的语义之后，可以推理得出 H2 这个命题为假，它们构成了矛盾关系(Contradiction)，又称阴性蕴含关系(Negative Textual Entailment)；对于 T3-H3，人们在获知 T3 的语义后，并不能以此为据判定命题 H3 的真假，因此，它们所构成的关系称为未知蕴含关系(Unknown Entailment)，又称中性关系(Neutral).

除非特别说明，本文中所提到的两个文本构成蕴含关系，指的都是由 T 可以推理得出 H 的阳性蕴含关系.

1.2.2　文本蕴含与其他文本间关系的区别与联系

首先，文本蕴含的研究范畴要和复述(Paraphrasing)进行区分. 复述，通常用来表示两个文本片段包含的相同的语义. 所以严格来讲，复述可以认为是一种语义上的对等(Textual Equivalence)关系，或者叫做双向蕴含关系(Bi-directional Textual Entailment). 而文本蕴含关系是单向推理关系. 如上例中，T1 可以推理得出 H1，但反之不可. 关于复述和蕴含的异同，Androutsopoulos 等人[12]作出了系统性的辨析.

T4：斑马是食草动物.
T5：野马是食草动物.

另外，文本蕴含的研究范畴要和文本相似(Text Similarity)有关研究进行区分. 文本相似，指的是一对文本包含的相似的语义. 比如，T4-T5 这对文本，其语义是相似的，都表示了某种动物对于更大范畴的归属关系；同时，如果使用编辑距离或其他相似度量进行考察，两句的文本相似度也较高. 但是它们并不构成蕴含关系，因为“斑马”和“野马”并不构语义上的包含关系，从而 T5 的语义并不能由 T4 推理得出. 事实上，文本相似度常常用作识别蕴含关系的重要特征[13-16]，但是相似的文本未必构成蕴含关系.

最后，文本蕴含关系并不是严格数学意义上的逻辑推理(Logical Inference)关系. 从文本蕴含的定义可以看出，判别 T-H 间是否构成文本蕴含关系，关键在于一般人类读到 T 之后，能否以 T 所包含的语义命题为依据，结合自身知识，判断 H 的语义命题的真伪. 尽管有些识别文本蕴含关系的方法[17-22]借鉴了逻辑推理的基本思想，但文本蕴含关系并不严格遵守数学逻辑推理原则，其判别过程也与数学逻辑推理过程不同.

综上，学者站在不同的角度去考察文本的相关属性，就得出了不同的文本间关系，它们既有区别，也互相联系，具体如表 1 所示.

表 1　常见文本间关系对比

文本间关系	特点	示例
文本蕴含	(1) 侧重于文本语义间的包含或推理属性； (2) 具有方向性； (3) 以人类常识作为推理依据.	丁磊 1997 年创办网易公司. 丁磊是网易公司的创办人.
文本相似	(1) 侧重于文本间的字面、结构或语义的相似属性； (2) 没有方向性； (3) 相似未必蕴含	斑马是食草动物. 野马是食草动物.
文本复述	(1) 侧重于同一语义的表述形式的丰富性； (2) 没有方向性； (3) 严格的文本复述等价于双向蕴含.	丁磊是网易公司的创办人. 丁磊创办了网易公司.
逻辑推理	(1) 侧重于推理过程的严谨性、完备性和正确性； (2) 具有方向性； (3) 以命题作为推理的基本单位； (4) 以数学定理或假设作为推理依据.	同一平面的两条直线没有交点. 这两条直线平行.

1.3　文本蕴含的基本问题

在自然语言处理领域，有 3 类围绕二元关系所展开的基本研究，分别是二元关系的识别、二元关系的抽取以及二元关系的生成. 所谓“识别”，是指给出一对可能构成某二元关系的文本对，要求机器对其关系是否成立给出判定；所谓“抽取”，就是要求机器能够自动地从大量自然语言文本中把构成该二元关系的成对文本片段抽取出来；所谓“生成”，是指给出一个文本片段，要求机器能够生成与之构成该二元关系的另外一方. 由此可见，“识别”是研究二元关系的第 1 步，其目的是为了“培养”机器对这种二元关系的认知鉴别能力；“抽取”则是第 2 步，目的是利用机器对该二元关系的鉴别能力，从自然语言文本中

获取大量的样本，积累知识；“生成”是第 3 步，做到这 1 步就可以认为机器已经掌握了该二元关系，能够举一反三，“灵活应用”了.

由定义 1 可知，文本蕴含关系也是一种二元关系，因此，对应地也有这 3 个基本问题，即文本蕴含关系的识别、文本蕴含知识的获取以及文本蕴含对的生成.

1.3.1 文本蕴含关系识别

识别文本蕴含关系是全部文本蕴含研究的基础. 下面给出给出其定义.

定义 2. 给定一对文本 T-H，要求机器对其是否构成蕴含关系做出判定，这样的问题称为识别文本蕴含关系(Recognizing Textual Entailment，RTE).

识别文本蕴含关系形式上是一种文本对分类问题. 其中最基本的是二元分类，即分为蕴含关系与非蕴含关系，此外，也可以把非蕴含关系进一步划分为矛盾关系和中性关系，形成多元分类. 另外，有学者单独研究矛盾关系，提出了矛盾检测问题(Contradiction Detection)[23].

识别文本蕴含关系本质上是一种基于语义进行推理的过程. 因此，其中有两个核心问题需要考虑——语义表示和推理机制. 这是一对相辅相成而又互相矛盾的问题. 一方面，在识别文本蕴含关系的过程中，语义的表示形式是为方便推理机制的执行而设计的，反过来，推理机制也能一定程度上弥补文本对 T-H 的语义表示上的鸿沟，因此，二者相互配合，缺一不可；另一方面，语义建模的鲁棒性和推理机制的严谨性却是一对不可兼得的矛盾. 自然语言处理技术中有一系列语言分析(Language Analysis)工具，诸如分词(Word Segmentation)、词性标注(POS Tagging)、句法分析(Parsing)、语义角色标注(Semantic Role Labeling)、形式化逻辑表示(Formal Logic Representation)等. 在这个序列中，自前至后各个语言分析工具对文本语义的刻画越来越精确：分词仅仅是把文本按单词进行切割，词性标注在此基础上增加了词性信息……形式化逻辑表示已经把语义表示成为精确的数学逻辑了. 越严谨的推理机制对语义表示的精确性要求就越高，例如，若能把文本对 T-H 用形式逻辑表示成为两个命题，就可以借用数学上严谨完备的机器证明工具进行推理；但如果仅用单词或词性来对语义建模，就只能应用单词重叠度、相似度或其他简易的启发式方法进行“模糊推理”了. 事实上，在上述序列中自前至后语言分析的难度在不断增加，同时靠后的语言分析工具也依赖前面的分析结果，由于错误级联效应，导致语义建模的正确性不断降低. 而即使应用严谨的推理机制也不能保证基于错误的语义表示所得到的推理结果的正确性. 故而在识别文本蕴含关系的实际应用中，需要有一个折衷取舍(trade-off)的考量. 语义表示与推理机制这两个核心问题的关系如图 1 所示(其中出现的推理机制将在 2.1 节中详细介绍).

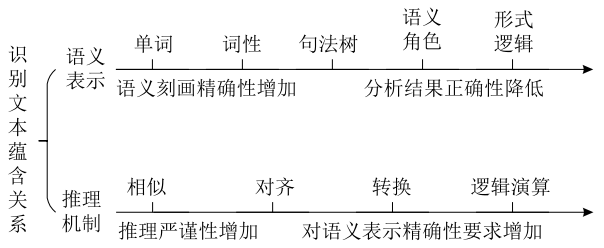

图 1 文本蕴含关系识别的核心问题

1.3.2 文本蕴含知识获取

由前面的讨论可知，对文本中蕴含现象的识别能力是获取蕴含知识的基础. 反过来，识别文本蕴含关系也离不开相关蕴含知识的积累，尤其是基于逻辑演算或转换的方法，其性能直接依赖于可应用的蕴含知识.

广义地讲，所谓蕴含知识就是对于识别文本蕴含关系有用的知识. 狭义地讲，蕴含知识是由 LHS(Left-hand Side)和 RHS(Right-hand Side)两部分组成的蕴含规则，即“LHS⇒RHS”. 例如，若已知“苹果⇒水果”，就可以得出“他吃了一个苹果. ⇒他吃了一个水果.”；若已知“X 购买了 Y⇒X 拥有 Y”，就可以得出“我买了一台电脑. ⇒我拥有一台电脑.”下面给出文本蕴含知识获取任务的定义：

定义 3. 给定一个文本集合 S，要求机器对其中构成文本蕴含关系的文本片段以“LHS⇒RHS”的形式抽取出来，该任务称为文本蕴含知识获取(Textual Entailment Knowledge Acquisition).

文本蕴含知识获取研究中也有两个核心问题需要考虑——知识表示和知识来源.

知识的表示形式是为了方便应用而设计的，蕴含知识根据是否含有变量可以划分为两类：单词及短语级别的蕴含知识(不含有变量，如“苹果⇒水果”)和模板级别的蕴含知识(含有变量，如“X 购买了 Y⇒X 拥有 Y”). 事实上，蕴含知识的应用场景往往是特定的，很少有放之四海而皆准的蕴含知识. 例如，“acquire”作为及物动词既有“购买”的意思，也有“学习”的意思，蕴含知识“X acquire Y⇒X purchase Y”在“AT&T acquire(收购) T-Mobile⇒AT&T

purchase T-Mobile”的上下文中成立，但在“Children acquire(习得)skills⇒Children purchase skills”的场景下中并不成立，因此如何对蕴含知识的应用场景进行建模是知识表示问题中需要考虑的地方.

蕴含知识的潜在来源有很多，例如词典、百科、新闻语料、普通互联网文本等等.按照是否有专家参与构建可以把知识源分为人工构建的资源和大规模语料两类，前者小而精，后者广而粗，针对不同的知识来源需要设计不同的知识获取方法.图 2 展示了文本蕴含知识获取的两个核心问题，2.2 节将从围绕这两个问题对当前蕴含知识获取研究取得的进展进行梳理.

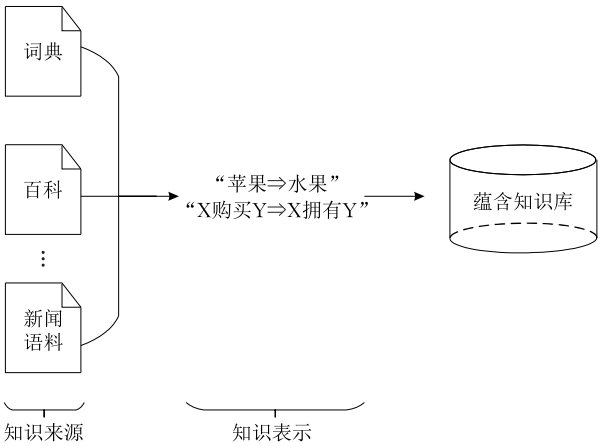

图 2　文本蕴含知识获取的核心问题

1.3.3　文本蕴含对的生成

定义 4.　给定一个文本片段 T 和蕴含知识库 D，要求机器根据 D 生成能够被 T 蕴含的文本片段 H，使 T⇒H 成立，该任务称为文本蕴含对的生成.

从定义 4 可以看出，文本蕴含对的生成任务实际上是在模拟人类根据自身掌握的知识(D)对给定线索(T)进行推理的过程.目前文本蕴含领域的研究主要集中在文本蕴含的关系识别和知识获取两个任务上，对文本蕴含对的生成方面研究较少[24-26].其原因大致有以下几个方面.

内因：一方面，H 的候选项个数随推理步数的增加呈指数级增长：假设蕴含知识库中的每个 LHS 平均对应 3 个不同的 RHS，那么经过一步推理可能产生 3 个一级候选项(RHS_1，RHS_2，RHS_3)，由于蕴含关系的传递性，则可能产生 9 个二级候选(RHS_{11}，RHS_{12}，…，RHS_{33})……3^N 个 N 级候选项.另一方面，推理的可靠性随推理步数的增加迅速降低：假设知识库中的每个推理规则的平均可靠度为 0.8，当可靠度低于 0.5 时认为推理不可靠，那么由于错误的级联效应，经过四次推理，其可靠度为 $0.8^4\approx0.4<0.5$ 就可以认为其正确性难以保证.因此，在研究文本蕴含对生成问题时，在推理广度和深度上都要进行有效剪枝.

外因：一方面，蕴含对的生成需要依赖蕴含识别技术所提供的推理机制和知识获取技术所积累的知识库，但是当前关系识别和知识获取的研究尚不够成熟，推理机制不够鲁棒，知识库中的推理规则的完备性和实用性也有所欠缺.另一方面，由于推理的发散性，文本蕴含对的生成技术缺少广泛的应用场景.目前已有学者[25]在对话系统领域进行了尝试.

1.3.4　识别、获取与生成的关系

作为文本蕴含研究领域的 3 个基本问题，文本蕴含的关系识别、知识获取与蕴含对生成 3 项研究彼此联系，相辅相成，构成了一个紧密结合的整体，其关系如图 3 所示.

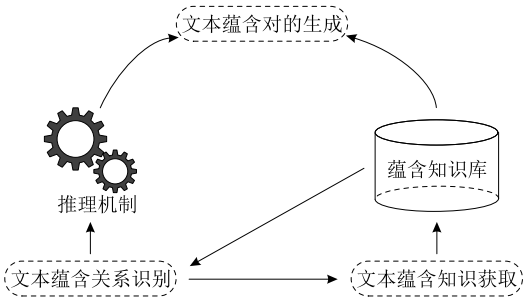

图 3　文本蕴含的基本问题及其关系

文本蕴含关系识别研究是文本蕴含有关研究的基石，培养了机器的对蕴含的识别能力，“输出”了推理机制；而文本蕴含知识的获取需要识别技术对自然语言文本中的蕴含知识进行识别，进而输出蕴含知识库；同时，蕴含知识库对某些基于转换或演算的识别研究提供了便利；而文本蕴含对的生成则需要推理机制和蕴含知识库共同为其提供动力.

1.4　章节安排

本节中的 1.1 节对文本蕴含的研究背景、应用场景、研究目的进行了阐述；1.2 节给出文本蕴含的定义，与其它文本间关系的区别与联系，界定其研究范畴；1.3 节指出文本蕴含研究的 3 个基本问题——关系识别、知识获取和蕴含对的生成；1.3.1 节指出关系识别的两个核心问题——语义表示与推理机制，并给出他们的制约关系；1.3.2 节指出知识获取的两个核心问题——知识表示与知识来源；1.3.3 节指出目前蕴含对生成相关研究进展缓慢的内因和外因；最后，1.3.4 节讨论识别、获取与生成的关系.

第 2 节对文本蕴含的研究进展进行归纳梳理.

首先,关系的识别研究是文本蕴含相关研究的基石;2.1 节将围绕推理机制对其研究脉络进行梳理,详细介绍各种机制的基本思想、演化关系及其可取与不足之处;其次,文本蕴含关系的识别离不开对蕴含知识的积累;2.2 节将围绕知识表示和知识来源这两个核心问题来讨论怎样挖掘蕴含知识,并指出不同知识获取方法的优缺点;最后,文本蕴含相关研究的蓬勃发展离不开近年来频繁举办的国际评测;2.3 节将简要介绍其中影响较大的几个国际评测和一些常用的数据集,并给出相应的评价指标和当前的最好成绩.

经过多年的发展,文本蕴含相关研究取得了一定的进展,但尚未完全达到成熟实用的水平.大数据时代的到来和近期深度学习理论与实践的发展为文本蕴含研究领域同时带来了机遇与挑战.第 3 节将立足当前研究现状,结合未来发展形势并对文本蕴含的发展前景进行展望.

2 研究现状分析

2.1 蕴含关系的识别方法

文本蕴含关系的识别研究是文本蕴含相关研究的基石,本节将对其研究脉络进行梳理,介绍不同方法的基本思想、演化关系及其可取与不足之处.

2.1.1 基于相似度的文本蕴含关系识别方法

构成蕴含关系的 T-H 对往往比较相似,如前例中的 T1-H1.因此,有人提出可以利用 T-H 对的相似程度来判断其是否构成蕴含关系.这就是基于相似度的文本蕴含关系识别方法的基本思想.

这类方法比较直观,在 RTE 研究领域的早期曾经一度是主流方法,但随着研究的不断深入,现在多把相似度作为判别模型的一个特征[27-30].

Jijkoun 等人[16]首先提出了基于词袋模型(Bag of Words)的文本蕴含关系识别方法.他的具体做法是首先把句子分词,通过词频对单词进行赋权,然后计算 Lin-相似度[31]和 WordNet[32]相似度,并以此为依据判断蕴含关系.

Adams[15]在 Jijkoun 的词袋模型基础上,创造性地提出了一种新的相似度量.该方法利用从 WordNet[32]中抽取出的词链来连接 T 和 H,并计算两者之间的编辑距离,最终结合其他特征使用决策树识别蕴含关系.类似地,张鹏等人[33]利用 FrameNet[34]中连接 T-H 的框架路径以及框架元素相似度来判断蕴含关系.

Mehdad 等人[14]进一步丰富了编辑距离算法,首先提出了基于句法树的相似度计算模型,其编辑操作定义在 T 和 H 的句法树节点上,但其基本操作较为简单.Heilman 等人[13]则在 Mehdad 等人的基础上定义了更加复杂的句法树编辑操作,包括对子节点、父节点、兄弟子树的增加、改动、删除等操作,最终不仅提高了文本蕴含关系的识别率、也在复述和问答系统的答案选择等任务上都取得了不错的成绩.

任函等人[35]提出了一种基于话题相似性的 RTE 方法.该方法认为存在蕴含关系的文本应当具有相似或相同话题.该方法利用知识话题模型(Knowledge-Based Topic Model)来计算语义相似度,并以此为依据判定蕴含关系.Saikh 等人[36]首次将机器翻译中用于评价系统译文和标准译文近似程度的指标当作 T-H 对的相似度量.

这些基于相似度的方法实现比较简单,同时方便设计各种相似度量,但是这种方法强行假设"相似即蕴含",导致大量语义相似但并非构成文本蕴含关系的实例被错误识别,例如 T4-T5.

基于相似度的识别方法的一般形式化表示如下:

$$F(T,H)=\begin{cases}1, & \text{若}\sum_i w_i sim_i(T,H)>\theta\\ 0, & \text{否则}\end{cases},$$

其中:函数 $F(T,H)$ 是蕴含关系的判定函数,若 $F(T,H)=1$,则判定 T⇒H,若 $F(T,H)=0$,则判定 T⇏H;θ 为阈值,取值为正;$sim_i(T,H)$ 是某种相似度量,w_i 为其权重.

2.1.2 基于对齐的文本蕴含关系识别方法

在基于相似度的识别方法基础上,演化出了基于对齐的识别方法.这类方法并不是直接使用相似度来判断蕴含关系,而是先把 T 和 H 中相似的部分找出来进行对齐,然后把对齐的方式和程度作为判断是否构成蕴含的依据.

De Marneffe 等人[23]首次把对齐和判断蕴含分成了两个步骤,并人工标注了部分对齐数据,然后应用机器学习方法学习对齐参数,实现了自动对齐辅助识别蕴含关系的方法.

Iftene[37]提出了非监督的自动化映射方法.他首先使用 DIRT[38]、WordNet[32]、VerbOcean[39]、Wikipedia 等外部知识库把 H 中的单词向 T 中的对应部分做映射,计算局部对齐程度(Local Fitness),归一化后换算成为全局对其程度(Global Fitness).如果存在多种映射方式,优先选择全局对齐程度最

大的映射方式.最终通过比较全局对齐程度是否超过阈值来判定该 T-H 对是否构成蕴含关系.在前人提出的基于单词的对齐方法的基础上,MacCartney 等人[40]提出了短语级别的对齐方法.Basak 等人[41]提出基于句法树的对齐方法.Sultan 等人[42]提出了基于混合特征的对齐方法.Noh 等人[43]提出了一种级联式多层次的对齐方法,可以灵活组合多种对齐模型综合判断蕴含关系.

基于对齐的方法在基于相似的方法基础上引入了对齐操作,从而能够着重比较 T-H 对中不相同的部分,这种方法优点在于其直观性,但是这些方法大多不能灵活处理 T-H 对中单词之间的复杂对齐方式,例如一对多,多对多,交叉对齐等.

基于对齐的识别方法的一般形式化表示如下:

$$F(T,H)=J(A_{opt}),$$

$$A_{opt}=\underset{\{(T_i\sim H_i)\}\in\Phi}{\arg\max}\,G(\{(T_i\sim H_i)\}),$$

$$G(\{T_i\sim H_i\})=\frac{L(\{T_i\sim H_i\})+\alpha M(\{T_i\sim H_i\})}{\#\{T_i\sim H_i\}},$$

$$L(\{T_i\sim H_i\})=\sum_i sim(T_i,H_i),$$

其中:函数 $F(T,H)$是蕴含关系的判定函数,若 $F(T,H)=1$,则判定 T⇒H,若 $F(T,H)=0$,则判定 T⇏H;$T_i\subset T$ 是文本 T 的一部分,$H_i\subset H$ 是文本 H 的一部分;$T_i\sim H_i$表示 T_i与 H_i互相对齐;$\{T_i\sim H_i\}$为 T-H 间某种可行的对齐方式;A_{opt}是最佳对齐方式;Φ是所有可能的对齐方式的全集;$J(\cdot)$是根据对齐方式判定蕴含关系的函数,当 T⇒H 时值 1,否则值为 0;$G(\{T_i\sim H_i\})$为全局对齐程度;$L(\{T_i\sim H_i\})$为局部对齐程度;$M(\{T_i\sim H_i\})$为全局对齐函数,通常$\{T_i,H_i\}$整体上对齐得越"规整",其取值越大;α为平衡系数,旨在平衡全局对齐函数和局部对齐程度;$\#\{T_i\sim H_i\}$为该对齐方式中一一对齐部分的个数,用于对全局对齐程度进行正则化;$sim(T_i,H_i)$为 T_i与 H_i的相似度.

2.1.3　基于逻辑演算的文本蕴含关系识别方法

蕴含关系实际上是一种语义推理关系,而数学界对命题逻辑证明问题已经有了比较成熟的方法与工具.因此,把逻辑演算的思想迁移到文本蕴含的识别上是非常自然的想法.基于逻辑演算的方法一般首先把 T-H 对和背景知识库中的事实编码为数学逻辑的表达式,比如一阶逻辑表达式(First Order Logic Expression),构成事实集合,然后应用逻辑推理规则来判断 H 的表达式是否可以由 T 的表达式和背景知识库所构成的事实集合所推出.

Hobbs 等人[44]首次应用溯因推理(Abductive Reasoning)的方法进行文本推理.溯因推理的基本思想是通过试图找到某个命题成立的原因的方式进行推理.他们构建了相当数量的常识和领域事实集合,对语义的逻辑化表示做了早期探索.

Raina 等人[17]利用句法依存关系将 T-H 对分别表示成为两组子命题的和取形式,其中子命题由句法依存树的节点转化而成,应用溯因推理机制尝试由 T 推出 H,并计算推理的代价,从而据此对 T-H 是否构成蕴含关系进行判别.

Moldovan 等人[18]实现的 COGEX 系统把 WordNet[32]释义以及 T-H 对表示成为逻辑命题,然后应用演绎法进行推理,最终应用到问答系统的答案验证(Answer Validation)中,对其性能提升了约 30%.演绎法是一种不断应用蕴含式(Implications,例如 LHS⇒RHS)的逻辑推理方法.简要地说,首先把命题 T 作为初始状态,如果当前状态匹配了某个蕴含式的前件(Antecedent, LHS),那么就可以应用这个蕴含式,把当前状态推导为蕴含式后件(Consequent,RHS)的形式;如果经过一系列的推导之后,最终得到了命题 H,那么判定该 T-H 对构成蕴含关系,这一系列推导就是形式化证明的过程;如果经过一系列推导,最终也未能推出 H,那么认为 T-H 不够成蕴含关系[45].另外,也可以利用反证法去推理,即由 H 的否命题去推导 T 的否命题.

Akhmatova[19]也做了类似的工作,并与其他方法进行了对比.他们实现了两个 RTE(Recognizing Textual Entailment)系统,第 1 个系统只使用简单的基于单词重叠率的特征,第 2 个使用了一阶逻辑演算和其他复杂特征,但实验结果表明这两个系统的准确率(Accuracy)相差无几.经过分析发现,尽管有部分 T-H 对可以通过逻辑演算得出,但是这种方法不能覆盖大部分 T-H 对,导致虽然系统识别的精确率较高但召回率较低;同时,在能够由 T 推导得出 H 的情况下,单词重叠率也很高,最终导致该方法性能未能领先基于单词重叠率的方法.

事实上,目前所积累的背景知识的覆盖范围是有限的,而且从自然语言到数学逻辑表达式的转化过程中也难免引入错误,如果一个 RTE 系统严格遵守逻辑证明原则,其召回率会非常低,例如 Bayer 等人[20]的实现的 System 1.为了避免这个问题,很多 RTE 系统都引入了一些松弛策略.例如 COGEX[18]系统允许忽略 T 或 H 中的一些子命题,而且在不影响最终置信度的前提下,也可以引入相对不可靠的推

理规则.而 Raina 等人[17]也引入了代价(Cost)机制.

基于逻辑演算的方法把数学界机器证明领域成熟的思想迁移到文本蕴含识别领域,具有一定的理论基础.但文字命题到逻辑表达形式的转化不够鲁棒,导致容错性较差;而背景知识的缺失往往使得推理链条中断,从而召回率偏低[19,21].

此外,由于基于逻辑演算的方法的数学严密性,可以用来处理一些其他方法很难解决的特殊文本蕴含问题.例如,数量蕴含(Quantity Entailment)[22]问题,即需要对 T-H 对中所提到的数字进行运算才能判断其蕴含关系的特定蕴含问题,如 T6-H6.Roy 等人[22]通过数学演算的方法解决了该问题,并为此专门设计了一种特殊的 QVR(Quantity-Value Representation)表达式.

T6:桌面上放着两本书.
H6:桌面上放着不到五本书.

基于逻辑演算的识别方法一般形式化表示如下:

$$F(T,H)=P(FOL(T)\cap FOL(D),FOL(H)),$$

其中:D 为蕴含知识库;函数 $F(T,H)$ 是蕴含关系的判定函数,若 $F(T,H)=1$,则判定 T⇒H,若 $F(T,H)=0$,则判定 T⇏H;$FOL(\cdot)$ 函数可以把文本参数转换为一阶逻辑表达式;函数 $P(\cdot,\cdot)$ 是数学证明函数,有两个参数,如果可以由第 1 个参数证明第 2 个参数,则输出 1,否则输出 0.

2.1.4 基于转换的文本蕴含关系识别方法

针对基于逻辑演算的方法的不足,有学者提出了基于转换的识别方法.这类方法采用了类似的"演算"思想,却抛弃了严格的数学逻辑表达式,转而利用语言分析技术,例如句法分析、语义角色标注等,将 T-H 对表示成为某种语言学表示形式,例如句法树、依存图等,并把背景知识表示成为推理规则,然后以这种表示形式进行推理.具体地讲,就是对 T 和 H 的表示形式依据推理规则做转换改写,争取向对方形式贴近,最后采用诸如判定子图同构的方法来判断 T-H 是否构成蕴含关系.

例如,Bar-Haim 等人[46-47]把 T-H 都表示成为句法树,然后依照规则对 T 的句法树进行改写,若最终能够改写成类似 H 的句法树形式,则判定蕴含关系成立;若应用了所有可能的改写规则,仍然不能在容许的范围内达成目标,则判定为非蕴含关系.类似地,Lien 等人[48]将 T-H 对表示成为语义图后进行改写.MacCartney 等人[49-50]则提出了基于 Natural Logic 的转换推理方法.该方法把所应用规则的序列作为特征,通过决策树来判定蕴含关系.

Stern 等人[51]在 Bar-Haim 等人的工作[47]基础上加入了与 Raina 等人[17]类似的代价(Cost)机制,而这些代价也是从训练集上学习而得到的.Tian 等人[52]沿用了 Bar-Haim 等人的思想,以 DCS 树(Dependency Compositional Semantics)的形式来进行推理.

基于转换的方法保留了基于逻辑演算方法的合理内核,同时由于不再要求把 T-H 转化为逻辑表达式,进而避免了因转化而引入的噪声.但是该类方法性能严重依赖转换规则,不完善的规则导致识别召回率降低,而错误的规则导致识别准确率降低.这些转换规则作为蕴含知识,有些来自于人工构建的知识库,也有一些来自大规模语料,2.3 节将对如何获取蕴含知识进行介绍.

基于转换的识别方法的一般形式化表示如下:

$$F(T,H)=SEARCH(REP(T),REP(D),REP(H)),$$

其中:D 为蕴含知识库;函数 $F(T,H)$ 是蕴含关系的判定函数,若 $F(T,H)=1$,则判定 T⇒H,若 $F(T,H)=0$,则判定 T⇏H;$REP(\cdot)$ 函数可以把文本参数转换为内在语义表示,如句法树、语义图等;函数 $SEARCH(\cdot,\cdot,\cdot)$ 是搜索函数,共有 3 个参数,如果可以利用第 2 个参数所代表的推理规则将第 1 个参数改写成为第 3 个参数的形式,则输出 1,否则输出 0.

2.1.5 基于混合模型的文本蕴含识别方法

针对前面所介绍的各类文本蕴含识别方法的优势与不足,有学者[53]提出了基于混合模型的 RTE 方法.该类方法把诸如 T-H 的词级别相似度(如单词重叠率、同义词、反义词等)、句法树相似度、句法树编辑距离、对齐程度、由 T 转换为 H 的代价等等混合在一起作为特征,送入分类器(如支持向量机等)进行分类的方法.

Zhang 等人[27]提出了融合了最小信息树(Minimum Information Tree)等特征的混合模型来解决中文文本蕴含识别问题.该方法将 T-H 中公共子序列在各自的句法树上合并成为一个节点,然后尝试合并相似的子树来简化句法树的结构,最终形成最小信息树.基于最小信息树的相似度是该模型中最有效的特征.最小信息树的优势在于它一定程度上避免了语言分析阶段(如中文分词、句法分析等)所引入的错误,从而提高了方法的鲁棒性.除了最小信息树这一句法特征,Zhang 等人还融合了单词重叠率、句子余弦相似度、知网[54]相似度、同义词林相似

度[55]、反义词、否定词等特征.

有很多针对中文的文本蕴含识别方法也可以归入这一类别. 如 Huang 等人[30]、李妍[56]、赵红燕等人[29]、盛雅琦等人[28]、刘茂福等人[57-59]、任函等人[60]、Zhang 等人[61]提出的一系列融合了 T-H 中同义词、反义词、单词重叠率、词性(POS)相似度、FrameNet[34]相似度、句法相似度、混合主题模型相似度[28]、事件语义相似度[58]、Word Embedding[61]等特征的分类器方法.

基于混合模型的识别方法的一般形式化表示如下：

$$F(T,H)=C(X),$$

$$X=(x_1^{T,H},x_2^{T,H},\cdots,x_N^{T,H}),$$

其中：函数 $F(T,H)$是蕴含关系的判定函数，若 $F(T,H)=1$，则判定 T⇒H，若 $F(T,H)=0$，则判定 T⇏H；映射 $C:\Re^N\rightarrow\{0,1\}$是分类器，分类为 T⇒H 时输出 1，否则输出 0，例如朴素贝叶斯、最大熵、支持向量机等；$\boldsymbol{X}$ 是 N 维实数向量；$x_i^{T,H}$表示 T-H 对的第 i 个特征.

2.1.6　基于深度神经网络的文本蕴含关系识别方法

随着深度学习理论研究的不断深入和实践的不断进展，近期有学者提出了一些基于深度神经网络的文本蕴含识别方法.

(1) 基于受限玻尔兹曼机的 RTE 方法

受限玻尔兹曼机(Restricted Boltzmann Machines，RBM)模型由 Smolensky[62]于 1986 年首次提出，它是一种学习训练集概率分布的随机生成神经网络模型. Hinton 等人[63]在 2006 年使用 RBM 进行降维，开启了深度学习时代. RBM 在特征学习、协同过滤以及主题建模等领域都有所应用[64]. Lyu 等人[65]将 RBM 模型应用到了识别文本蕴含关系领域. 他们首先建立了一个 RBM 模型来学习 T-H 对的联合表示，并用该模型重建 T-H 对，然后利用重建误差(Reconstruction Error)作为特征对该 T-H 对是否存在文本蕴含关系进行判定. 作者认为，重建误差越小，T-H 对的共同语义越多，从而越有可能存在蕴含关系. 除了 RTE 领域，该模型也可用于复述检测(Paraphrase Detection)问题.

(2) 基于递归神经网络的 RTE 方法

递归神经网络(Recursive neural network)由 Goller 等人[66]所提出. Socher 等人[67]首先将递归神经网络应用到句法分析领域. 其后，Socher 等人[68]与 Irsoy 等人[69]又将递归神经网络用于情感分析领域. 与 RTE 任务类似，情感分析也是一个文本分类的任务，只不过它只对一个文本 T 的情感极性进行判断，而 RTE 任务需要对一个文本对 T-H 的蕴含关系进行识别. 递归神经网络在这些领域的成功证明了其对自然语言文本的语义具有建模的能力.

随后，Bowman 等人[70]首次使用递归神经网络来识别文本蕴含关系，并取得了不错的成绩. 该方法使用递归神经网络对 T 和 H 分别进行建模，得到两个向量，然后再将两个向量送入隐含层进行比较，最终通过 Softmax 分类器进行分类. 递归神经网络的优势在于其可以充分利用句法信息，但其性能也比较依赖句法分析结果.

(3) 基于卷积神经网络的 RTE 方法

卷积神经网络(Convolutional Neural Networks，CNN)是一类基于空间上卷积操作的神经网络模型. Kalchbrenner 等人[71]和 Kim[72]尝试将其用于自然语言处理领域，对句子进行建模. Yin 等人[73]提出了基于 Attention 机制的卷积神经网络来处理 RTE 问题. 由于识别蕴含关系需要考虑一对文本 T-H，因此，可以在对其中一个文本(如 T)进行建模时参照另一个文本(H)的信息，这种参照机制就是 Attention 机制. 该方法通过在卷积操作中加入 Attention Matrix 的方法来实现这种参照.

此外，普通的卷积神经网络不能有效地捕获句法信息，而这些信息是判别文本蕴含关系重要特征. Mou 等人[74-75]为此提出了基于树结构的卷积神经网络(Tree-Based Convolutional Neural Networks，TBCNN). 该网络将依存句法子树作为卷积操作的应用对象，构成子树特征提取器，它能够一次性提取父节点和其所有子节点间的依存关系信息. 该方法分别使用两个 TBCNN 来对 T-H 建模，使用拼接、求差、按维度相乘等启发式特征构造向量代表 T-H 对的语义信息，然后送入 Softmax 分类器进行分类. 对比之前的 CNN 类句子模型，该方法额外利用了依存关系的类型信息，从而提高了识别准确率.

(4) 基于 LSTM 类神经网络的 RTE 方法

LSTM(Long Short-Term Memory)神经网络[76]是一类由门(Gate)控制的循环神经网络(Recurrent Neural Networks)，它擅长对一维序列进行建模，并且通过门机制解决了一般循环神经网络中的梯度消失(Vanishing gradient)问题. Sundermeyer 等人[77]提出基于 LSTM 的句子模型. 具体地，按照从左到右的顺序，依次将句子中的每个词送入 LSTM 神经网络，输入完毕后，用 LSTM 的最终状态输出来表示这个句子. Bowman 等人[78]首次将 LSTM 句子模型带入 RTE 领域.

Rocktäschel 等人[79]针对 RTE 任务改进了 LSTM 句子模型.他分别使用两个 LSTM 先后对 T 和 H 进行建模,并用第 1 个 LSTM 的最终状态来初始化第 2 个 LSTM.在建模的过程中也引入了 Attention 机制,即在第 2 个 LSTM 处理 H 时参照了 T 的信息.此外,Rocktäschel 等人还提出了 Word-by-word Attention 机制,即在第 2 个 LSTM 读取 H 中每个词的时候,都引入第一个 LSTM 处理 T 所输出的信息,进一步提升了模型性能.通过分析 Attention 向量(在 Word-by-word Attention 机制中为矩阵)发现,该机制能较好地关注到 T-H 对中的语义对应部分,并且实现了软对齐,对传统的基于对齐的 RTE 方法[23,37,49]进行了突破.类似的,Liu 等人[80]提出了基于带有 Attention 机制的双向 LSTM(BiLSTM)模型来识别文本蕴含关系.

Wang 等人[81]在 Rocktäschel 等人工作的基础上,提出了 mLSTM(matching-LSTM)模型,重点关注 T-H 对中各部分的匹配情况.该方法把对 H 建模的 LSTM 模型的输出与对 T 建模的 LSTM 模型产生的 Attention 向量进行拼接,作为输入送入 mLSTM,并把 mLSTM 的最终输出送入 Softmax 分类器进行分类.此外,该方法在 T 中增加了一个特殊单词 NULL,如果 H 中当前单词与 T 中的所有正常单词均不构成匹配,mLSTM 就会把它与 NULL 进行匹配,相当于增加了悬空对齐方式,完善了对齐模型.通过对 mLSTM 的忘记门(forget gate)进行分析发现,mLSTM 倾向于记住不匹配(mismatches)信息而忘记匹配信息.如果最后没有记住不匹配信息,就认为 T-H 的语义是相匹配的,从而判定蕴含关系成立,否则将其判定为冲突或中性关系.通过对 mLSTM 的输入门(input gate)的分析发现,重要的实词(content words)会被 mLSTM 重视,而停用词往往被忽视.mLSTM 的机制符合人类直觉,设计的比较巧妙.

Cheng 等人[82]从人类的阅读习惯中获得灵感,提出了 LSTMN(Long Short-Term Memory-Networks)模型来识别蕴含关系.与传统 LSTM 模型相比,LSTMN 使用记忆带(memory tape)而不是记忆槽(Memory Cell)来记忆以往的状态(state)和输出(output),解决了之前 LSTM 类方法记忆压缩导致信息损失的问题.并且该模型在 LSTM 内部添加了一个 Attention 层来实现序列内部的参照.实验表明,内部 Attention 层能够捕获句内依存关系.对于双序列建模问题,Cheng 等人设计了 Shallow Attention Fusion 和 Deep Attention Fusion 两种参照机制.类似 Wang 等人的工作[81],Cheng 等人在 T-H 两句的末尾分别加上了一个特殊单词(EOS).

基于神经网络的方法是当今 RTE 最前沿(State-of-the-art)方法.该类方法具有识别准确率高、鲁棒性强、领域移植性好等优点,但神经网络收敛速度往往较慢,模型优化也需要一定的技巧,此外模型的参数众多,导致需要大量的训练语料才能充分训练.

基于深度神经网络的识别方法的一般形式化表示如下:

$F(T,H)=\mathrm{Softmax}(NN(SM(T),SM(H)))$,

其中,函数 $F(T,H)$是蕴含关系的判定函数,若 $F(T,H)=1$,则判定 T⇒H,若 $F(T,H)=0$,则判定 T⇏H;Softmax(·)是 Softmax 分类器,若 T⇒H,输出 1,否则输出 0;$SM(\cdot)$是句子模型函数,负责把文本 T 或 H 转换成为向量表示形式,可以是 CNN、LSTM 等神经网络句子模型;$NN(\cdot)$是用于拼接及比较向量的非线性神经网络函数.

2.1.7 各类文本蕴含关系识别方法对比

经过十余年的发展,学界涌现了一大批思路迥异但行之有效的蕴含关系识别方法,前文对这些方法的发展脉络进行了梳理,表 2 将对各类方法的可取之处与不足之处进行总结.

表 2 蕴含关系识别方法对比

推理机制	可取之处	不足之处
相似度	(1)方法直观,实现简单; (2)方便应用各种相似度量.	强行假设“相似即蕴含”,存在大量被错误识别的负例.
对齐	在相似的基础上,引入对齐操作,着重比较不同部分,思路符合人类直觉.	对齐方式不够灵活,不能处理一对多、多对多等.
逻辑演算	(1)借用数学上成熟的“机器证明”思想,具有一定的理论基础; (2)能够解决诸如数量蕴含等其他方法难以判断的问题.	(1)由于文本到逻辑命题的转换鲁棒性不足,导致容错性较差; (2)背景知识缺乏导致推理链条中断,召回率低.
转换	(1)保留了基于逻辑演算方法的合理内核; (2)不要求转换为逻辑式,一定程度上提高了鲁棒性.	方法性能严重依赖转换规则:不完善的规则导致召回率低,错误的规则导致识别准确率低.
混合模型	融合了多种推理机制的特征,综合性强,从而适用性较广	(1)方法不够直观; (2)需要较多的训练语料
深度神经网络	(1)连续化向量表示,克服了特征稀疏问题; (2)神经网络能一定程度上模拟了人脑的思维机制; (3)识别准确率高; (2)鲁棒性强; (3)领域可移植性好.	(1)模型参数众多; (2)学习收敛速度慢; (3)需要大量的训练语料.

2.2 蕴含知识的获取方法

文本蕴含关系识别研究离不开相关蕴含知识的积累.尤其是基于逻辑演算或转换的方法,其性能直

接依赖于可应用的蕴含知识. 由 1.3.2 节可知，知识来源与知识表示是蕴含知识获取研究的两个核心问题. 其中，蕴含知识来源可以分为人工构建的资源和大规模语料两种；蕴含知识按照表示方式不同可划分为两类——单词及短语级别的蕴含知识和模板级别的蕴含知识. 下面将围绕这两个核心问题对蕴含知识获取方法进行介绍.

2.2.1　从手工构建的资源中获取蕴含知识的方法

(1) 单词及短语级别的蕴含知识获取

人工构建的资源主要有词典和百科两类. 词典中的词项一般由单词及其释义构成. 而单词的释义中，有可能含有该单词的同义词、上位词等蕴含知识.

WordNet[32] 释义精炼、形式规范，是在 RTE 系统中广泛用到的机读词典(Machine Readable Dictionary). Moldovan 等人[83]首先尝试将 WordNet 中单词释义转化成为逻辑表达式. Kouylekov 等人[84]利用 WordNet[32] 中词汇的同义词、上下位词等关系获取单词及短语级的蕴含知识. Pazienza 等人[85]利用 WordNet 中单词间关系、VerbNet[86] 中动词间框架关系以及 WordNet 与 VerbNet 的映射关系来获取蕴含知识.

维基百科等在线百科是的蕴含知识的另一大来源. Kouylekov 等人[87]用基于 LSA 的词汇相似度方法从维基百科中获取单词及短语级蕴含知识；此外，百科数据的结构性较强，有些结构暗示了蕴含关系，例如，在维基百科的标题中有很多用于解释的括号结构“W(V)”，这种结构其实暗示着 W⇒V 这一蕴含关系，如由“The Siren(Musical)”可得出“The Siren⇒Musical”. Shnarch 等人[88]手工总结了维基百科中类似的蕴含模式，然后通过模式匹配的方法获取单词及短语级蕴含知识.

(2) 模板级别的蕴含知识获取

FrameNet[34] 是一个按照语义框架进行组织的英语词典资源. 其中的每个框架对应一种事件类型，包括属于该框架的谓词、论元以及例句. Aharon 等人[89]提出了一种利用 FramNet 框架间关系从中例句中获取蕴含模板(如 cure X⇒X recover)的方法.

由于人工构建的知识库覆盖了比较重要的词汇关系，并且结构性较强，因此从其中获取到的蕴含知识的准确率较高. 但是知识库中词汇相对固定，导致该类方法所获取的蕴含知识规模比较受限.

2.2.2　从大规模语料中获取蕴含知识的方法

除了词典等人工构建的资源，大规模新闻语料或 Web 检索结果中同样包含丰富的蕴含知识，因此有学者提出了从大规模语料中获取文本蕴含知识的方法.

(1) 单词及短语级别的蕴含知识获取

Harris[90] 提出了“分布假设”的思想，即具有相似上下文的单词或短语的语义是相似的. 而语义相似的单词或短语往往具有蕴含关系.

Lin[91] 基于分布假设思想提出了 Lin-相似度，并用它在大规模语料中获取单词及短语级蕴含知识. Lin-相似度的特点是，如果两个词 W-V 的上下文重合率较高，这两个词的 Lin-相似度也会比较高. 如果某两个词 W-V 的 Lin-相似度超过某个阈值，就认为它们之间存在蕴含关系. 该方法从语料中获取了大量的单词及短语级蕴含知识，但其中难免有一些噪声. 一些上下文相似却不具备蕴含关系的反义词，如“good-bad”，会混在所获取的蕴含知识中. 此外，该方法没有给出两个词之间的蕴含方向.

对此，Geffet 等人[92]提出了一种确定单词间蕴含方向的方法. 其基本思想是，对给定的两个词 W-V，如果在所有 V 的上下文中都可以把 V 替换成为 W，就认为 W 比 V 的适用范围更广，从而认为 W⇒V. 此外，Zanzotto 等人[93]提出了基于“选择偏好(selectional preference)”的动词间蕴含方向判定方法. 所谓选择偏好是指动词的语义角色类型，语义角色类型较多的动词蕴含语义角色较少的动词. Kotlerman 等人[94] 进而提出了有方向的相似度(Directional Similarity Measure)来确定蕴含方向.

(2) 模板级别的蕴含知识获取

Lin 等人[38]沿用分布假设思想，又提出了一种模板级蕴含知识获取方法 DIRT. 对于一个模板来说，例如“X buy Y”，可以把它的上下文定义为所有可以填充槽 X 和 Y 的单词集合. 两模板的相似度可以用槽 X 的 Lin-相似度[91] 与槽 Y 的 Lin-相似度的几何平均值表示. 和基于分布假设的单词及短语级蕴含知识获取方法类似，DIRT 获取的知识中经常包含构成反义的模板对，如“X solves Y”-“X worsens Y”. 另外，该方法没有给出蕴含方向. 但实际上，该方法所获取到的模板对中，只有大约 20%～25%是双向蕴含的复述模板对[95]，而大量其它模板对只存在单向蕴含关系，如果在应用模板时不加以区分会对系统带来不确定性.

Szpektor 等人[96] 从另一个角度提出了基于 bootstrapping 的 TEASE 方法. 该方法可以根据人工提供的种子模板迭代地从语料中获取蕴含知识. 具体地，首先用种子模板在搜索引擎中检索得到一

个包含该模板的句子集合,并在这个集合中把所有可填充该模板的单词提取出来;然后用所提取到的单词进行检索得到新的句子集合,并从中把所有这些单词能填充的模板抽取出来;之后再用这些新得到的模板进行迭代检索……

经过过滤[97]后的模板都可被认为与原始种子模板具有蕴含关系.此外,也可以应用蕴含关系的传递性来丰富蕴含模板[98].基于 Bootstrapping 的方法能够获得非常丰富的蕴含模板资源,但该方法需要较多的人工干预,也没有给出模板之间蕴含方向.

Bhagat 等人[99]提出一种称为 LEDIR(LEarning Directionality of Inference Rules)的方法判定一对模板 W-V 的蕴含方向.与 Zanzotto 等人[93]提出的选择偏好方法类似,LEDIR 方法利用 W 和 V 的填充词的语义类别数量来判断蕴含关系.

另外,模板的使用情况也应当考虑上下文.例如,"X acquire Y⇒X purchase Y"在"AT&T acquire(收购)T-Mobile⇒AT&T purchase T-Mobile"的上下文中成立,但在"Children acquire(习得)skills⇒Children purchase skills"中并不成立.Séaghdha[100]、Dinu 等人[101]和 Melamud 等人[102]利用 LDA(Latent Dirichlet Allocation,一种主题模型)对模板的上下文按主题聚类,通过在不同主题下分别计算相似度的方法,一定程度上解决了蕴含模板的上下文敏感问题.

与人工构建的资源相比,大规模语料中的蕴含知识表现形式更加丰富,但这些语料往往缺乏良好的结构,导致从中获取的蕴含知识的准确率较低.

2.2.3 各类文本蕴含知识获取方法对比

表 3 对前面介绍的各类蕴含知识获取方法进行了总结对比.

表 3 蕴含知识获取方法对比

方法	知识来源	可取之处	不足之处
从人工构建资源中获取	WordNet VerbNet FrameNet Wikipedia	(1) 词典覆盖了的词汇上下位关系; (2) 百科结构性较强; (3) 从中获取到的蕴含知识的准确率较高.	(1) 语料构建成本高; (2) 词汇相对固定,导致该类方法所获取的蕴含知识规模比较受限.
基于分布假设	大规模新闻语料; Web 检索结果	(1) 具有相似上下文的短语/模板往往具有蕴含关系; (2) 方法简单; (3) 语料易收集.	(1) 语料结构性差,获取的知识准确率低; (2) 新闻语料表达方式单一,获取的知识覆盖度低; (3) Web 检索结果的规范性和一致性差,导致蕴含规则的应用性能较低.
基于 Bootstrapping	大规模语料库; Web 检索结果	大规模语料库中表述方式的多样性使获取的蕴含模板规模较大	需要较多的人工干预(需人工制定种子模板、过滤规则)

2.3 与文本蕴含有关的国际评测

近年来,国际上多家机构举办了多届与文本蕴含有关的国际评测,为各个方法提供了横向比较的平台;同时每一届评测都会发布新的数据集,从而极大地促进了文本蕴含相关研究的进展.本章节将简要介绍其中影响较大的几个国际评测和一些常用的数据集,并给出相应的评价指标和当前的最好成绩.

2.3.1 PASCAL RTE 评测

RTE(Recognizing Textual Entailment)评测是最早举办的面向文本蕴含的国际评测,自 2005 年起由欧洲的 PASCAL Network 资助举办,从 2008 年的 RTE-4 开始变为 NIST(National Institute of Standards and Technology,美国国家标准与技术研究所)所组织的 TAC(Text Analysis Conference)评测的一部分,至 2011 年共连续举办了 7 届[103-109],在文本蕴含关系识别的研究初期,按年度举办的 RTE 评测吸引了大量学者从事相关研究.

RTE 的评测语料库中的 T-H 对主要来自需要处理文本蕴含关系的 4 个应用场景——问答系统(Question Answering)、关系抽取(Relation Extraction)、信息检索(Information Retrieval)和文摘(Summarization),其主要语言为英语.RTE-1 至 RTE-5 是经典的蕴含关系识别任务,其语料规模、平衡性,最佳成绩等详见表 4.在前 3 届中只考察蕴含和非蕴含两种关系,从 RTE-4 开始将非蕴含关系进一步划分为矛盾和未知两种,并开始引入长文本,进而考察语篇(Discourse)间的蕴含关系.RTE-6 开始引入搜索任务,即要求系统在语料库中寻找能够蕴含 H 的文本 T.此后,RTE-8 评测与 Semeval-2013 的 SRA(Student Response Analysis)评测合并,其形式与前 7 届 RTE 评测差异较大,将在 2.3.5 节单独介绍.

经典 RTE 评测的主要指标是 T-H 对的分类准确度(*Accuracy*),即分类正确的比率:

$$Accuracy \triangleq \frac{1}{\#pairs}\sum_{i} 1[\hat{y}_i = y_i],$$

其中:$\#pairs$ 是 T-H 对的总数;$\hat{y}_i$ 是系统对第 i 个 T-H 对的关系判定标签,y_i 是其正确答案;$1[\cdot]$ 是指示函数,当其参数为真时值为 1,否则为 0.

RTE-1～RTE-3 的语料可从 PASCAL 的网站①上获取;RTE-4～RTE-7 的语料可从 TAC 的网站②上获取.经过历届的 RTE 评测,识别文本蕴含

① http://pascallin.ecs.soton.ac.uk/
② http://www.nist.gov/tac/data/index.html

关系的研究从无到有，诞生了一系列思路迥异的识别方法.

2.3.2　CLEF AVE 评测

CLEF(Cross Language Evaluation Forum，后改名为 Conference and Labs of the Evaluation Forum)，是一个致力于对跨语言信息处理系统进行评测的机构.该机构曾于 2006 年至 2008 年连续举办了 3 届 AVE(Answer Validation Exercise，答案验证)评测[110-112].

AVE 评测的目标是开发一个评估系统，要求其能自动地判断给定问答系统的作答是否正确. AVE 语料来自 CLEF QA 评测任务，由多种语言构成，包括巴斯克语、保加利亚语、德语、英语、西班牙语、法语、意大利语、荷兰语、葡萄牙语、罗马尼亚语和希腊语等.下面给出答案验证(AVE)任务的定义.

定义 5.　给定三元组〈问题 Q，答案 A，支持文本 ST〉，要求系统根据 ST 判断 A 是否是 Q 的正确答案，这样的任务称为答案验证任务，简称 AVE 任务.

由定义 5 可知，AVE 任务也是一个文本分类问题，其评价指标为 *F-score* 等.

$$F\text{-}score=\frac{2\times recall\times precision}{recall+precision},$$

$$precision=\frac{\text{是正确答案且判定正确的样本数}}{\text{判定为正确答案的样本数}},$$

$$recall=\frac{\text{是正确答案且判定正确的样本数}}{\text{是正确答案的样本数}}.$$

有些参与评测的系统使用了 RTE 技术[113-114].这些 RTE 系统中的 H 来自于由问题 Q 和系统作答案 A 合并得到的陈述句，T 则来自对应的支持文档 ST.一般情况下，如果系统作答正确，T-H 间必然存在蕴含关系.这就是通过识别 T-H 间的蕴含关系来对答案进行评估的基本思想.故而，AVE 评测可以看作是对 RTE 技术的间接评测.参与评测的系统用到了基于相似度、转换、逻辑演算的推理机制，评测结果表明应用 RTE 技术能够有效地提升问答系统的性能[110].

2.3.3　CLEF QA4MRE 评测

CLEF 在 2011 年～2013 年连续举办了 3 次 QA4MRE(Question Answering for Machine Reading)评测[115].这个评测的任务形式类似于高考阅读理解题目，所涉及的语言有阿拉伯语、保加利亚语、英语、西班牙语和罗马尼亚语等.其任务定义如下.

定义 6.　给定一篇文章和与之相关的背景知识语料库，同时给出几个问题及若干候选项，每个问题的候选项中至多有一个是正确的答案，系统可以从候选项中选择一项作为答案或选择不作答，这样的任务称为机器阅读问答任务，简称 QA4MRE 任务.

要完成 QA4MRE 任务一般需要两个步骤：首先要在文章或语料库中找到每个候选答案 A 的支持文本 ST，然后按照与 AVE 任务类似的方法对候选答案三元组〈Q，A，ST〉进行验证.其中第 1 个步骤需要信息检索技术，第 2 个步骤需要答案验证技术.因此，QA4MRE 评测也是一种对 RTE 技术的间接评测[116-117].

评测的指标是 *accuracy* 和 *c*@1，定义如下：

$$accuracy=\frac{n_R+n_{UR}}{n},$$

$$c@1=\frac{1}{n}\left(n_R+n_U\frac{n_R}{n}\right),$$

其中：n 是所有问题总数；n_R 是正确回答的问题数；n_U 是未回答的问题数；n_{UR} 为候选项中没有正确答案，同时系统也选择不作答的问题数.

通过 QA4MRE 评测，证明了 RTE 技术可以用于提升问答系统的性能，同时也促进了 RTE 技术的发展.此外，这些评价指标的设计非常巧妙，尤其是加入了对没有正确选项的问题的考察，促使参与评测的系统先尝试去真正“理解”文章，再去考虑作答.

2.3.4　NTCIR RITE 评测

NTCIR(NII Testbeds and Community for Information access Research)是日本国立情报学研究所资助的一系列研讨会，致力于推动信息处理技术相关研究的发展.2011 年的 NTCIR-9 开始举办 RITE(Recognizing Inference in Text)评测，参评系统需要识别给定文本对的蕴含、复述或矛盾关系.数据集语言有英语、日语、简体中文和繁体中文[30]等.至 2014 年 NTCIR-11，RITE 评测共举办了 3 届[118-120]，数据集可从 NTCIR 网站①上获取.

每年的评测任务略有调整，但基本任务形式如下：

二元分类任务(Binary-Class)，给定一个文本对 T-H，需要判断 H 是否可由 T 推理得出；

多元分类任务(Multi-Class)，给定文本对 T-H，需要把它们的关系分为四类：T 单向蕴含 H

① http://research.nii.ac.jp/ntcir/index-en.html

(Forward Entailment,简记为 F)、T 与 H 双向蕴含(Bi-directional Entailment,简记为 B)、T 与 H 矛盾(Contradiction,简记为C)、T与H独立(Independence,简记为 I).

RITE 任务的评价指标是 *Accuracy*,其与经典的RTE任务相同,详见2.3.1节.表4中列举了RITE任务中与中文相关的常用数据集,并介绍了其语料规模、平衡性、最佳成绩等信息.

2.3.5 SemEval 相关评测

SemEval(Semantic Evaluation)是一个致力于促进各类语义分析方法发展的研讨会. SemEval的前身是 Senseval(Word Sense Disambiguation Evaluation),该评测前期主要关注词义消歧任务,后来逐渐加入其他语义分析任务并改为按年度举办. 2010 年起开始引入有关文本蕴含的评测任务. SemEval-2010 举办了 PETE(Parser Evaluation using Textual Entailments)评测[11],要求利用文本蕴含技术评价句法分析结果. SemEval-2012、SemEval-2013 举办了两届CLTE(Cross-Lingual Textual Entailment)评测[121-122],要求识别来自不同语言的文本对 T-H 是否构成蕴含关系. CLTE 评测要求对 T-H 间关系进行四元分类:前向蕴含、后向蕴含、双向蕴含、非蕴含. SemEval-2013 增设 SRA(Student Response Analysis)评测[9],SRA 评测要求参测系统对学生的作业进行打分. 由于学生的作业与标准答案之间的蕴含关系可以指示学生答案的完善程度,因此这届 SRA 评测又被 SemEval 称为 RTE-8 评测. SemEval-2014 的task 1[123]所衍生的SICK(Sentences Involving Compositional Knowledge)语料库①,包含了约 10000 个有三元蕴含标注(蕴含、冲突、中性)的英语 T-H 对,为近期识别文本蕴含研究提供了标准评测语料[124],表 4 介绍了其语料规模、平衡性、最佳成绩等信息. SICK 语料库的评价指标是 *Accuracy*,与经典的 RTE 任务相同,详见 2.3.1 节.

表 4 常用数据集对比

数据集名称	语言	发布时间	数据集规模(T-H 对数量)		T-H 对关系	类别分布	最好成绩	
			训练集/开发集	测试集			*Accuracy*	推理机制
RTE-1	英语	2004	567	800	蕴含+非蕴含	50%:50%	0.519[19]	逻辑演算
RTE-2	英语	2005	800	800	蕴含+非蕴含	50%:50%	0.6262[15]	相似度
RTE-3	英语	2006	800	800	蕴含+非蕴含	50%:50%	0.670[107]	相似度
RTE-4	英语	2008	未发布[106]	1000	蕴含+中性+冲突	50%:35%:15%	0.614[125]	混合模型
RTE-5	英语	2009	600	600	蕴含+中性+冲突	50%:35%:15%	0.6833[126]	对齐
RITE-2-CS-BC	汉语(简体)	2012	814	781	蕴含+非蕴含	950:645	0.7465[127]	对齐
RITE-2-CS-MC	汉语(简体)	2012	814	781	B+F+C+I	304:646:252:393	0.6108[127]	对齐
RITE-2-CT-BC	汉语(繁体)	2012	1321	881	蕴含+非蕴含	1195:1007	0.6776[128]	混合模型
RITE-2-CT-MC	汉语(繁体)	2012	1321	881	B+F+C+I	413:872:368:549	0.5664[129]	混合模型
SICK	英语	2014	4934	4906	蕴含+中性+冲突	2821:5595:1424	0.769[70]	RNTN
SNLI	英语	2015	560152	10000	蕴含+中性+冲突	190113:189218:189702	0.863[82]	LSTMN

表 5 SNLI 数据集上模型对比

模型	模型参数数量	*Accuracy*
混合模型[78]	—	0.782
LSTM[130]	约 3000000 个	0.806
Tree-based CNN[74]	约 3500000 个	0.821
LSTM+attention[79]	约 252000 个	0.835
Bi-LSTM[80]	约 2000000 个	0.833
Bi-LSTM+attention[80]	约 2800000 个	0.850
mLSTM+attention[81]	约 1900000 个	0.861
LSTMN+attention[82]	约 3400000 个	0.863

2.3.6 SNLI 语料

SNLI(Stanford Natural Language Inference)[78]是斯坦福大学自然语言处理研究小组所发布的用于评测识别文本蕴含关系方法的语料库②. 该语料库规模远超之前所有语料,具有约 570000 个具有三元蕴含标注的英语 T-H 句对③,并且其类别分布相对均衡. 该语料库为需要大规模训练语料的神经网络类 RTE 模型提供了数据保证[74,79,81-82]. 由于该语料库的以上特点,渐渐在较新的研究工作[74,78-79,81-82]中成为标准评测集. 表 4 将其与以往的常用的经典 RTE 语料进行了对比. 表 5 对比了在 SNLI 语料上训练的最新识别模型. SNLI 语料库的

① SICK 语料库获取地址 http://clic.cimec.unitn.it/composes/sick.html.

② SNLI 语料库获取地址 http://nlp.stanford.edu/projects/snli/.

③ 除蕴含、冲突、中性外,SNLI 语料中还存在少量没有标签的样本.

评价指标是 *Accuracy*，与经典的 RTE 任务相同，详见 2.3.1 节.

2.3.7　常用数据集对比与当前最好成绩

本节通过表格的形式在数据集规模、类别分布等维度上对文本蕴含关系识别研究中常用的中英文数据集进行了对比，并且给出了在每个数据集上取得的最好成绩和对应的推理机制. 其中 SNLI 数据集语料规模大、类别分布平衡，是目前最优标准测试集，表 5 对其上展开的最新研究进展进行了总结.

关于数据集规模，早期发布的 RTE 系列英文语料和 RITE 系列中文语料中 T-H 对的数量一般都在 2000 对以下，规模比较有限；而近期发布的 SICK 和 SNLI 语料，其中的 T-H 对数量在 10 000 对以上.

由表 4 可知，在小数据集上，人工构建特征的经典方法，例如基于相似、对齐或逻辑演算的推理机制所取得的效果比较好；而在大数据集上基于深度神经网络的识别方法取得的成绩较好. 由 2.1.6 节的讨论可知，神经网络类方法能够自动学习 T 和 H 的表示向量，不需要人工构建特征，但是对训练样本的数量有比较高的要求，这里再次印证了这一点.

由表 5 可知，神经网络类模型需要学习的参数量巨大，几乎都在百万级，但由于训练样本充足，其识别性能已经超过了经典的混合模型.

3　未来研究展望

经过众多学者多年来的不断耕耘，在文本蕴含关系的识别和蕴含知识的获取两个领域涌现了不少思路迥异但行之有效的方法. 大数据时代的到来以及近期深度学习研究的蓬勃发展，为文本蕴含研究带来了丰富的知识来源和有力的科研工具. 可以预见，如何有效地利用这些便利促进文本蕴含研究将成为未来的研究热点. 本节立足当前研究形势，提出了几个未来研究方向以及一些理论上可行的研究思路.

3.1　大数据带来的机遇与挑战

随着“大数据”时代的到来，人们所积累的自然语言文本越来越多，这给文本蕴含相关研究带来诸多机遇与挑战.

3.1.1　丰富的知识来源和潜在训练语料

文本蕴含关系广泛存在于海量的自然语言文本中，其表现形式的多样性超过了所有以往人工构建的知识库或语料. 如果能有效地表示并提取这些丰富多样的文本蕴含知识，那么将很大程度上解决当前基于逻辑演算或转换的 RTE 方法中由于背景知识缺乏而导致推理链条中断的问题，进而提高这类方法的鲁棒性. 此外，神经网络类 RTE 方法模型复杂、参数众多，导致所需要的训练集规模非常大，而这些海量文本便成为了潜在的训练语料. 因此，大数据所带来的海量自然语言文本是能促进文本蕴含有关研究的宝藏.

那么，如何利用大数据来获取更多的蕴含知识以及潜在训练语料将是 1 个值得研究的课题. 笔者认为至少可以从以下 3 个出发点着手：

(1) 文本自身的弱标注线索

除人工构建的资源之外，自然语言文本大多是非结构化的，并且没有显式地标注出其中的蕴含关系，但是文本自身仍然存在一些指示蕴含关系的线索，例如：

关联词“那么”、“因此”有可能连接了一些存在蕴含关系的文本；

文本所提及事件发生时间的先后，也有可能指示了蕴含关系；

句式“X 是一种 Y”或“X，一种 Y”，可能暗示 X 蕴含 Y……

以往数据规模较小，导致这些弱标注线索分布比较稀疏，能匹配的蕴含现象非常有限，从而导致这些线索价值较低. 但是大数据时代带来了海量的自然语言文本，使得即使一些之前认为分布比较稀疏的弱标注线索仍然会对应大量的实例，因此可以通过在海量自然语言文本中检索弱标注线索的方法获取丰富多样的蕴含知识和潜在训练语料.

(2) 文本载体的弱结构信息

尽管海量自然语言文本是非结构化的，但是文本的载体或多或少会具有一些结构性信息，例如：

新闻或电子邮件至少有标题和正文两部分组成，而标题一般是正文的凝练概括，一般是可以由正文推理得出的. 这种弱结构信息就指示了两个文本片段——标题和正文——之间的蕴含关系；

Twitter 中的 Hash Tag 一般会按照话题事件进行组织，而描述这些话题的 Tweets 是带有时间信息的，而时间上的先后性有可能就暗示了两个 Tweets 之间的蕴含关系……

因此，可以利用文本载体的弱结构信息来获取蕴含知识和潜在训练语料.

(3) 蕴含关系的传递性

若 $O \Rightarrow P$，$P \Rightarrow Q$，那么 $O \Rightarrow Q$. 这就是蕴含关系

的传递性.抽象地讲,传递性是蕴含关系的固有属性之一,如果给我们一个蕴含知识库,我们就可以利用其中蕴含知识的传递性几乎不断地创造新的蕴含知识.但是在实际应用中,蕴含知识 O⇒P 或 P⇒Q 也有一定可能性不成立,由于错误的级联效应,O⇒Q 成立的可能性比 O⇒P 和 P⇒Q 都要低,而多次应用已有的蕴含知识所生成的新蕴含知识的正确性以往是难以验证的.其原因在于这样产生的新蕴含知识在一般语料中很难出现.但是大数据时代的到来为验证这些蕴含知识提供了可能.因此,通过大数据来验证蕴含关系的传递性也是一种可行的知识获取方式.

总之,大数据的规模效应是一根非常有用的杠杆,可以把诸如弱标注信息、弱结构信息以及蕴含关系的传递特性当作支点,来撬动大数据中的蕴含知识宝藏.

3.1.2 海量知识获取过程中的矛盾检测与消除

事实上,由于噪声等诸多因素的影响,由不规则的大数据得到的蕴含知识可能是互相矛盾的,如果最终获得的蕴含知识库不能做到逻辑自洽(Logical Consistency),那么将给相关应用带来不确定性.因此,海量知识获取过程中的矛盾检测与消除也是一个值得研究的课题.

对于这一问题,笔者认为可以从知识表示和知识来源两个角度考虑:

(1)细化知识表示

事实上,大多数蕴含知识都有自己的应用场景,很少有能放之四海而皆准的蕴含知识;因此,很多表面上看可能构成矛盾的蕴含知识可能只是由于其中一方脱离了自身的应用场景.目前蕴含知识的表示形式对应用场景的建模还不够精确是导致产生矛盾的重要原因.因此,可以尝试细化蕴含知识的应用场景或上下文来对其进行区分,从而尽量避免矛盾.

(2)对知识来源的置信度进行评估

尽管我们可以尽量细化蕴含知识的应用场景,但是如果有些知识来源本身就含有事实性错误,那么无论我们怎么增强对场景的精确建模能力,也不能避免矛盾的发生.这时,就可以考虑对知识来源的置信度进行评估,从而就可以对来自于该知识源的蕴含知识进行评分,当要使用的蕴含知识发生矛盾时,就可以采用舍弃评分较低的蕴含知识的方法来保证逻辑自洽性.同时,这种方法一定程度上也能保持应用的灵活性.

3.2 深度神经网络带来的机遇与挑战

3.2.1 基于深度学习的文本蕴含研究将成为热点

近年来,深度学习理论和实践不断深入发展,在语音识别和图像处理等领域获得了巨大的成功[131-132].由于神经网络是对人脑思考推理过程在抽象意义上的模拟,因此通过神经网络来处理文本蕴含关系理论上是可行的.同时,神经网络方法通过词向量(Word Embedding)的方式表示语义,解决了以往研究中语义表示稀疏所带来的问题[133].而深度学习方法解决了传统神经网络模型容易收敛到局部最优值所导致的训练困难等问题.因此,利用深度神经网络解决文本蕴含问题是水到渠成的.

目前在文本蕴含知识获取领域还没有基于深度神经网络的方法,但从 2015 年起开始陆续有学者把一些深度神经网络模型应用到文本蕴含关系识别领域,2.1.6 节已经对这些方法分门别类地进行了归纳总结.在撰写本文初稿的过程中,又有 3 种[82,130,134]新的基于深度神经网络的识别方法被提出,足可见深度学习在识别文本蕴含关系领域的热度.在不引入人工特征的情况下,这些模型基本都超过了以往的经典方法.此外,2016 年 6 月开始有学者[26]将深度神经网络用于文本蕴含对的生成任务上,该方法使用 SNLI 数据集来训练带有 Attention 机制的 LSTM 编码-解码网络,并直接使用该网络生成蕴含对,从而突破了生成任务必须要由知识库驱动的限制.可以预见,基于深度神经网络模型的文本蕴含研究将成为当前乃至未来的研究热点.

3.2.2 深度神经网络模型中值得研究的问题

基于深度神经网络的方法并非完美.首先,深度神经网络的参数规模非常大,例如,用于识别文本蕴含关系的 Tree-based CNN 模型[74]有大约 350 万个参数需要学习.这导致这类方法的性能非常依赖训练集的规模,在训练语料不足的情况下,神经网络模型的性能甚至低于使用人工定义特征的经典方法[78].而具有蕴含关系的海量的自然语言文本是未经标注的,如何利用未标注的海量文本来弥补训练集的不足从而有效训练深度神经网络是一个值得研究的课题.对此,3.1.1 节已经进行了讨论.

此外,当前深度神经网络模型的输入只是标注了蕴含关系的 T-H 对,与人类的推理过程相比,这类模型缺乏背景知识,有可能导致在某些情况下判断失误.而学者在研究经典方法时,积累了大量的蕴含知识[38,96,99].因此,如何把积累的蕴含知识导入深

度神经网络也是将来一个值得研究的课题.对此,笔者认为,结合已有的蕴含知识利用蕴含对生成技术来生成新的训练语料是一种可行的办法.

4 结　论

文本蕴含关系是广泛分布于自然语言文本中的单向推理关系,文本蕴含可以辅助其他自然语言处理任务的进行,并且具有丰富的应用场景,因此文本蕴含相关研究是自然语言处理领域的一项基础性研究.

本文首先界定了文本蕴含研究的范畴.作为一种二元关系,文本蕴含有 3 个基本研究任务——关系识别、知识获取和蕴含对生成.其中,关系识别有两个核心问题,即语义表示与推理机制;知识获取也有两个核心问题,即知识表示与知识来源;蕴含对生成研究进展缓慢有其内因和外因.

本文围绕语义表示与推理机制这两个核心问题梳理了关系识别的研究进展,围绕知识表示与知识来源梳理了知识获取的研究进展,并指出了各类方法的可取之处与不足之处.文本蕴含研究的进展离不开相关国际评测,本文也对这些国际评测和数据集进行了归纳总结.

大数据时代的到来和深度学习理论的不断发展,为文本蕴含相关研究提供了丰富的知识来源和有力的研究工具,同时也带来了许多崭新的研究课题.本文立足当前研究形势,展望了未来研究方向,并从理论上探讨了其可行性.

致　谢　《计算机学报》编辑部和各位审稿老师提出了宝贵意见,在此表示感谢!

参　考　文　献

[1] Dagan I, Glickman O. Probabilistic textual entailment: Generic applied modeling of language variability//Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining. Grenoble, France, 2004: 26-29

[2] Wang Bao-Xing, Zheng De-Quan, Wang Xiao-Xue, et al. Multiple-choice question answering based on textual entailment. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 134-140(in Chinese)
(王宝鑫, 郑德权, 王晓雪等. 基于文本蕴含的选择类问题解答技术研究. 北京大学学报(自然科学版), 2016, 52(1): 134-140)

[3] Bhaskar P, Banerjee S, Pakray P, et al. A hybrid question answering system for Multiple Choice Question (MCQ)//Proceedings of the Question Answering for Machine Reading Evaluation(QA4MRE) at Conference and Labs of the Evaluation Forum. Valencia, Spain, 2013: 1-18

[4] Iftene A, Gînsca A-L, Moruz M A, et al. Enhancing a question answering system with textual entailment for machine reading evaluation//Proceedings of the Conference and Labs of the Evaluation Forum (Online Working Notes/Labs/Workshop). Rome, Italy, 2012: 1-12

[5] Harabagiu S, Hickl A. Methods for using textual entailment in open-domain question answering//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Sydney, Australia, 2006: 905-912

[6] Romano L, Kouylekov M, Szpektor I, et al. Investigating a generic paraphrase-based approach for relation extraction//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006: 409-416

[7] Harabagiu S, Hickl A, Lacatusu F. Satisfying information needs with multi-document summaries. Information Processing & Management, 2007, 43(6): 1619-1642

[8] Padó S, Cer D, Galley M, et al. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. Machine Translation, 2009, 23(2-3): 181-193

[9] Dzikovska M O, Nielsen R D, Brew C, et al. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge//Proceedings of the 1st Joint Conference on Lexical and Computational Semantics. Atlanta, USA, 2013: 263-274

[10] Nielsen R D, Ward W, Martin J H. Recognizing entailment in intelligent tutoring systems. Natural Language Engineering, 2009, 15(4): 479-501

[11] Yuret D, Han A, Turgut Z. Semeval-2010 task 12: Parser evaluation using textual entailments//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden, 2010: 51-56

[12] Androutsopoulos I, Malakasiotis P. A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research, 2010, 38(1): 135-187

[13] Heilman M, Smith N A. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Los Angeles, USA, 2010: 1011-1019

[14] Mehdad M, Matteo N, Elena C, et al. EDITS: An open source framework for recognizing textual entailment//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2009: 169-178

[15] Adams R. Textual entailment through extended lexical overlap //Proceedings of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment. Venice, Italy, 2006: 128-133

[16] Jijkoun V, de Rijke M. Recognizing textual entailment using lexical similarity//Proceedings of the 1st PASCAL Challenge Workshop. Southampton, UK, 2005: 73-76

[17] Raina R, Ng A Y, Manning C D. Robust textual inference via learning and abductive reasoning//Proceedings of the National Conference on Artificial Intelligence (AAAI). Pittsburgh, USA, 2005: 1099-1105

[18] Moldovan D, Clark C, Harabagiu S, et al. Cogex: A logic prover for question answering//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canada, 2003: 87-93

[19] Akhmatova E. Textual entailment resolution via atomic propositions//Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. Southampton, UK, 2005: 61-64

[20] Bayer S, Burger J, Ferro L, et al. MITRE's submissions to the EU pascal RTE challenge//Proceedings of the Pattern Analysis, Statistical Modelling, and Computational Learning (PASCAL) Challenges Workshop on Recognising Textual Entailment. Southampton, UK, 2005: 44

[21] Bos J. Is there place for logic in recognizing textual entailment? Linguistic Issues in Language Technology, 2013, 9(3): 1-18

[22] Roy S, Vieira T, Roth D. Reasoning about quantities in natural language. Transactions of the Association for Computational Linguistics, 2015, 3: 1-13

[23] De Marneffe M-C, Rafferty A N, Manning C D. Finding contradictions in text//Proceedings of the ACL: HLT, Association for Computational Linguistics. Columbus, USA, 2008: 1039-1047

[24] Nevěilová Z. Paraphrase and textual entailment generation// Proceedings of the International Conference on Text, Speech, and Dialogue. Brno, Czech Republic, 2014: 293-300

[25] Jia J. The generation of textual entailment with NLML in an intelligent dialogue system for language learning CSIEC// Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering. Piscataway, USA, 2008: 194-201

[26] Kolesnyk V, Rocktäschel T, Riedel S. Generating natural language inference chains. arXiv preprint arXiv: 1606.01404, 2016

[27] Zhang Z, Yao D, Chen S, et al. Chinese textual entailment recognition based on syntactic tree clipping//Proceedings of the 13th China National Conference on Computational Linguistics and the 2nd International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. Wuhan, China, 2014: 83-94

[28] Sheng Ya-Qi, Zhang Han, Lv Chen, Ji Dong-Hong. Textual entailment recognition based on mixed topic model. Computing Engineering, 2015, 41(5): 180-184(in Chinese)
(盛雅琦, 张晗, 吕晨, 姬东鸿. 基于混合主题模型的文本蕴涵识别. 计算机工程, 2015, 41(5): 180-184)

[29] Zhao Hong-Yan, Liu Peng, Li Ru, Wang Zhi-Qiang. Recognizing textual entailment based on multi-features. Journal of Chinese Information Processing, 2014, 28(2): 109-115(in Chinese)
(赵红燕, 刘鹏, 李茹, 王智强. 多特征文本蕴涵识别研究. 中文信息学报, 2014, 28(2): 109-115)

[30] Huang W-J, Liu C-L. Exploring lexical, syntactic, and semantic features for Chinese textual entailment in NTCIR RITE evaluation tasks. arXiv preprint arXiv: 1504.02150, 2015

[31] Lin D. Extracting collocations from text corpora//Proceedings of the 1st Workshop on Computational Terminology. Montreal, Canada, 1998: 57-63

[32] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[33] Zhang Peng, Li Guo-Chen, Li Ru, et al. Recognize text entailment based on FrameNet relations. Journal of Chinese Information Processing, 2012, 26(2): 46-50(in Chinses)
(张鹏, 李国臣, 李茹等. 基于 FrameNet 框架关系的文本蕴含识别. 中文信息学报, 2012, 26(2): 46-50)

[34] Baker C F, Fillmore C J, Lowe J B. The berkeley FrameNet project//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. Montreal, Canada, 1998: 86-90

[35] Ren Han, Sheng Ya-Qi, Feng Wen-He, et al. Recognizing textual entailment based on knowledge topic models. Journal of Chinese Information Processing, 2015, 29(6): 119-126(in Chinese)
(任函, 盛雅琦, 冯文贺等. 基于知识话题模型的文本蕴涵识别. 中文信息学报, 2015, 29(6): 119-126)

[36] Saikh T, Naskar S K, Giri C, et al. Textual entailment using different similarity metrics//Proceedings of the International Conference on Intelligent Text Processing and Computational. Cario, Egypt, 2015: 491-501

[37] Iftene A. UAIC Participation at RTE4//Proceedings of the 1st Text Analysis Conference. Gaithersburg, USA, 2008: 1-10

[38] Lin D, Pantel P. DIRT@ SBT@ discovery of inference rules from text//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 323-328

[39] Chklovski T, Pantel P. VerbOcean: Mining the Web for fine-grained semantic verb relations//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 33-40

[40] MacCartney B, Galley M, Manning C D. A phrase-based alignment model for natural language inference//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Waikiki, USA, 2008: 802-811

[41] Basak R, Kumar Naskar S, Pakray P, et al. Recognizing textual entailment by soft dependency tree matching. Computacion Y Sistemas, 2015, 19(4): 6257-6259

[42] Sultan M A, Bethard S, Sumner T. Feature-rich two-stage logistic regression for monolingual alignment//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 949-959

[43] Noh T-G, Padó S, Shwartz V, et al. Multi-level alignments as an extensible representation basis for textual entailment algorithms//Proceedings of the Joint Conference on Lexical and Computational Semantics. Denver, USA, 2015: 193

[44] Hobbs J R, Stickel M, Martin P, et al. Interpretation as abduction//Proceedings of the 26th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, 1988: 95-103

[45] Toledo A. Semantic modeling of textual entailment: Proof-based annotation in a compositional framework. Netherlands: LOT, 2015

[46] Bar-Haim R, Berant J, Dagan I, et al. Efficient semantic deduction and approximate matching over compact parse forests //Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2008:1-10

[47] Bar-Haim R, Dagan I, Greental I, et al. Semantic inference at the lexical-syntactic level//Proceedings of the National Conference on Artificial Intelligence. Vancouver, Canada, 2007: 871-876

[48] Lien E, Kouylekov M. Semantic parsing for textual entailment //Proceedings of the International Conference on Parsing Technologies. Bilbao, Spain, 2015: 40

[49] MacCartney B, Manning C D. Natural logic and natural language inference. Computing Meaning, 2014, 4: 129-147

[50] MacCartney B, Manning C D. Natural logic for textual inference//Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Stroudsburg, USA, 2007: 193-200

[51] Stern A, Dagan I. A confidence model for syntactically-motivated entailment proofs//Proceedings of the 8th International Conference on "Recent Advances in Natural Language Processing". Hissar, Bulgaria, 2011: 455-462

[52] Tian R, Miyao Y, Takuya M. Logical inference on dependency-based compositional semantics//Proceedings of the Conference of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 79-89

[53] Sharma N, Sharma R, Biswas K K. Recognizing textual entailment using dependency analysis and machine learning// Proceedings of the NAACL-HLT 2015 Student Research Workshop (SRW). Denver, USA, 2015: 147

[54] Dong Zhen-Dong, Dong Qiang, Hao Chang-Ling. Theoretical findings of HowNet. Journal of Chinese Information Processing, 2007, 21(4): 3-9(in Chinese)
(董振东，董强，郝长伶. 知网的理论发现. 中文信息学报，2007，21(4)：3-9)

[55] Jiu-Le T, Wei Z. Words similarity algorithm based on tongyici cilin in semantic web adaptive learning system. Journal of Jilin University (Information Science Edition), 2010, 6(10): 602-608

[56] Li Yan. Multi-Features Based Textual Entailment Recognition in Chinese [M. S. dissertation]. Wuhan University of Science and Technology, Wuhan, 2013(in Chinese)
(李妍. 基于多特征的中文文本蕴涵识别方法[硕士学位论文]. 武汉科技大学，武汉，2013)

[57] Liu M, Guo Y, Nie L. Recognizing entailment in Chinese texts with feature combination//Proceedings of the 2015 International Conference on Asian Language Processing (IALP). Suzhou, China, 2015: 82-85

[58] Liu Mao-Fu, Li Yan, Ji Dong-Hong. Event semantic feature based Chinese textual entailment recognition. Journal of Chinese Information Processing, 2013, 27(5): 129-136(in Chinese)
(刘茂福，李妍，姬东鸿. 基于事件语义特征的中文文本蕴含识别. 中文信息学报，2013，27(5)：129-136)

[59] Liu M, Zhang L, Hu H, et al. A classification model for semantic entailment recognition with feature combination. Neurocomputing, 2016, 208: 127-135

[60] Ren Han, Wan Jing, Wu Hong-Miao, Feng Wen-He. A co-training based approach to recognizing textual entailment. Journal of Chinese Information Processing, 2014, 28(6): 114-119(in Chinese)
(任函，万菁，吴泓缈，冯文贺. 基于协同训练的文本蕴含识别. 中文信息学报，2014，28(6)：114-119)

[61] Zhang Z, Yao D, Pang Y, et al. Chinese textual entailment recognition enhanced with word embedding//Proceedings of the 14th China National Conference on Computational Linguistics. Guangzhou, China, 2015: 89-100

[62] Smolensky P. Information Processing in Dynamical Systems: Foundations of Harmony Theory. Cambridge, USA: MIT Press, 1986: 194-281

[63] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507

[64] Hinton G E, Salakhutdinov R R. Replicated softmax: An undirected topic model//Proceedings of the Conference on Neural Information Processing Systems. Vancouver, Canada, 2009: 1607-1614

[65] Lyu C, Lu Y, Ji D, et al. Deep learning for textual entailment recognition//Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI). Vietri sul Mare, Italy, 2015: 154-161

[66] Goller C, Kuchler A. Learning task-dependent distributed representations by backpropagation through structure//Proceedings of the IEEE International Conference on Neural Networks. Washington, USA, 1996: 347-352

[67] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks//Proceedings of the 28th International Conference on Machine Learning (ICML-11). Bellevue, USA, 2011: 129-136

[68] Socher R, Perelygin A, Wu J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, USA, 2013: 1631-1642

[69] Irsoy O, Cardie C. Deep recursive neural networks for compositionality in language. Advances in Neural Information Processing Systems, 2014, 3: 2096-2104

[70] Bowman S R, Potts C, Manning C D. Recursive neural networks can learn logical semantics. arXiv: 1406.1827, 2014

[71] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. arXiv preprint arXiv: 1404.2188, 2014

[72] Kim Y. Convolutional neural networks for sentence classification. arXiv preprint arXiv: 1408.5882, 2014

[73] Yin W, Schütze H, Xiang B, et al. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. arXiv: 1512.05193, 2015

[74] Mou L, Rui M, Li G, et al. Recognizing entailment and contradiction by tree-based convolution. arXiv: 1512.08422, 2015

[75] Mou L, Men R, Li G, et al. Natural language inference by treebased convolution and heuristic matching//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 130-136

[76] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[77] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. Interspeech, 2012, 31: 601-608

[78] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 632-642

[79] Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about Entailment with Neural Attention. arXiv: 1509.06664, 2015

[80] Liu Y, Sun C, Lin L, et al. Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv: 1605.09090, 2016

[81] Wang S, Jiang J. Learning natural language inference with LSTM. arXiv preprint arXiv: 1512.08849, 2015

[82] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading. arXiv preprint arXiv: 1601.06733, 2016

[83] Moldovan D I, Rus V. Logic form transformation of WordNet and its applicability to question answering//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Toulouse, France, 2001: 402-409

[84] Kouylekov M, Magnini B. Building a large-scale repository of textual entailment rules//Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 2006: 2347-2440

[85] Pazienza M T, Pennacchiotti M, Zanzotto F M. Mixing WordNet, VerbNet and propbank for studying verb relations //Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy, 2006: 1372-1377

[86] Schuler K K. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon[Ph.D. dissertation]. University of Pennsylvania, Philadelphia, 2005

[87] Kouylekov M, Mehdad Y, Negri M. Mining Wikipedia for large-scale repositories of context-sensitive entailment rules //Proceedings of the International Conference on Language Resources and Evaluation. Valletta, Malta, 2010: 3550-3553

[88] Shnarch E, Barak L, Dagan I. Extracting lexical reference rules from Wikipedia//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, USA, 2009: 450-458

[89] Aharon R B, Szpektor I, Dagan I. Generating entailment rules from FrameNet//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 241-246

[90] Harris Z S. Distributional structure. Word, 1954, 10(2-3): 146-162

[91] Lin D. Automatic retrieval and clustering of similar words//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2. Montreal, Canada, 1998: 768-774

[92] Geffet M, Dagan I. The distributional inclusion hypotheses and lexical entailment//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, USA, 2005: 107-114

[93] Zanzotto F M, Pennacchiotti M, Pazienza M T. Discovering asymmetric entailment relations between verbs using selectional preferences//Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia, 2006: 849-856

[94] Kotlerman L, Dagan I, Szpektor I, et al. Directional distributional similarity for lexical inference. Natural Language Engineering, 2010, 16(4): 359-389

[95] Szpektor I, Shnarch E, Dagan I. Instance-based evaluation of entailment rule acquisition//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, 2007: 456-463

[96] Szpektor I, Tanev H, Dagan D, et al. Scaling web-based acquisition of entailment relations//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004: 41-48

[97] Ravichandran D, Hovy E. Learning surface text patterns for a question answering system//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, 2002: 41-47

[98] Kloetzer J, Torisawa K, Hashimoto C, et al. Large-scale acquisition of entailment pattern pairs by exploiting transitivity //Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1649-1655

[99] Bhagat R, Pantel P, Hovy E H, et al. LEDIR: An unsupervised algorithm for learning directionality of inference rules //Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic, 2007: 161-170

[100] Séaghdha D O. Latent variable models of selectional preference//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 2010: 435-444

[101] Dinu G, Lapata M. Topic models for meaning similarity in context//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Stroudsburg, USA, 2010: 250-258

[102] Melamud O, Berant J, Dagan I, et al. A two level model for context sensitive inference rules//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 1331-1340

[103] Bentivogli L, Clark P, Dagan I, et al. The seventh pascal recognizing textual entailment challenge//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2011: 1-16

[104] Bentivogli L, Clark P, Dagan I, et al. The sixth PASCAL recognizing textual entailment challenge//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2010: 1-18

[105] Bentivogli L, Dagan I, Dang H T, et al. The fifth PASCAL recognizing textual entailment challenge//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2009: 1-15

[106] Giampiccolo D, Dang H T, Magnini B, et al. The fourth pascal recognizing textual entailment challenge//Proceedings of the 1st Text Analysis Conference. Gaithersburg, USA, 2008: 1-11

[107] Giampiccolo D, Magnini B, Dagan I, et al. The third pascal recognizing textual entailment challenge//Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Prague, Czech Republic, 2007: 1-9

[108] Bar-Haim R, Dagan I, Dolan B, et al. The second pascal recognising textual entailment challenge//Proceedings of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment. Venice, Italy, 2006: 1-9

[109] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge. Machine Learning Challenges, 2006, 3944: 177-190

[110] Rodrigo Á, Peñas A, Verdejo F. Overview of the answer validation exercise 2008//Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Aarhus, Denmark, 2009: 296-313

[111] Peñas A, Rodrigo Á, Verdejo F. Overview of the answer validation exercise 2007//Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Budapest, Hungary, 2008: 237-248

[112] Penas A, Rodrigo A, Sama V, et al. Overview of the answer validation exercise 2006//Proceedings of the Workshop of the Cross-Language Evaluation Forum. Alicante, Spain, 2007: 257-257

[113] Pakray P, Gelbukh A, Bandyopadhyay S. Answer validation using textual entailment//Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing. Tokyo, Japan, 2011: 353-364

[114] Wang R, Neumann G. DFKI-LT at AVE 2007: Using recognizing textual entailment for answer validation// Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages. Budapest, Hungary, 2007: 1-6

[115] Peñas A, Hovy E, Forner P, et al. QA4MRE 2011-2013: Overview of question answering for machine reading evaluation //Proceedings of the 4th International Conference of the CLEF Initiative. Valencia, Spain, 2013: 303-320

[116] Clark P, Harrison P, Yao X. An entailment-based approach to the QA4MRE challenge//Proceedings of the Conference and Labs of the Evaluation. Rome, Italy, 2012: 1-15

[117] Pakray P, Bhaskar P, Banerjee S, et al. A hybrid question answering system based on information retrieval and answer validation//Proceedings of the Conference and Labs of the Evaluation. Amsterdam, Netherlands, 2011: 1-16

[118] Shima H, Kanayama H, Lee C-W, et al. Overview of ntcir-9 rite: Recognizing inference in text//Proceedings of the 9th NII Test Collection for Information Retrieval Workshop. Tokyo, Japan, 2011: 291-301

[119] Watanabe Y, Miyao Y, Mizuno J, et al. Overview of the recognizing inference in text (RITE-2) at NTCIR-10// Proceedings of the 10th NII Test Collection for Information Retrieval Workshop. Tokyo, Japan, 2013: 385-404

[120] Matsuyoshi S, Miyao Y, Shibata T, et al. Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) task//Proceedings of the 10th NII Test Collection for Information Retrieval Workshop. Tokyo, Japan, 2014: 223-232

[121] Negri M, Marchetti A, Mehdad Y, et al. Semeval-2013 Task 8: Cross-lingual textual entailment for content synchronization// Proceedings of the International Workshop on Semantic Evaluation. Atlanta, USA, 2013: 25-33

[122] Negri M, Marchetti A, Mehdad Y, et al. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization //Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation. Montreal, Canada, 2012: 399-407

[123] Marelli M, Bentivogli L, Baroni M, et al. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment//Proceedings of the International Workshop on Semantic Evaluation. Dublin, Airland, 2014: 1-8

[124] Marelli M, Menini S, Baroni M, et al. A SICK cure for the evaluation of compositional distributional semantic models// Proceedings of the Ninth International Conference on Language Resources and Evaluation. Reykjavik, Iceland, 2014: 216-223

[125] Wang R, Neumann G. An divide-and-conquer strategy for recognizing textual entailment//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2008: 1-7

[126] Iftene A, Moruz M-A. Uaic participation at RTE5//Proceedings of the Text Analysis Conference. Gaithersburg, USA, 2009: 1-10

[127] Wang X, Zhao H, Lu B-L. BCMI-NLP labeled-alignment-based entailment system for NTCIR-10 RITE-2 task// Proceedings of the 10th NTCIR Evaluation of Information Access Techologies. Tokyo, Japan, 2013: 474-478

[128] Shih C-W, Liu C, Lee C-W, et al. IASL RITE system at NTCIR-10//Proceedings of the 10th NTCIR Evaluation of Information Access Techologies. Tokyo, Japan, 2013: 425-429

[129] Lin C-J, Tu Y-C. The description of the NTOU RITE system in NTCIR-10//Proceedings of the 10th NTCIR Evaluation of Information Access Techologies. Tokyo, Japan, 2013: 495-498

[130] Bowman S R, Gauthier J, Rastogi A, et al. A fast unified model for parsing and sentence understanding. arXiv preprint arXiv: 1603.06021, 2016

[131] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[132] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82-97

[133] Balkir E, Kartsaklis D, Sadrzadeh M. Sentence entailment in compositional distributional semantics. arXiv preprint arXiv: 1512.04419, 2015

[134] Parikh A P, Täckström O, Das D, et al. A decomposable attention model for natural language iInference. arXiv preprint arXiv: 1606.01933, 2016

GUO Mao-Sheng, born in 1991, Ph. D. candidate. His research interests focus on textual entailment.

ZHANG Yu, born in 1972, Ph. D., professor, M. S. supervisor. His research interests include question answering and personalized information retrieval.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include social computing, information retrieval and natural language processing.

Background

Research on textual entailment is a fundamental study in the field of natural language processing. It has a variety of applications, such as relation retrieval, question answering, multi-document summarization and machine translation. Many theories, models and methods of recognizing textual entailment and knowledge acquisition have been proposed and extensively studied. Although many achievements have been made in these areas, new problems are continually proposed and new challenges emerge. Especially, the arrival of big data era and the development of deep learning theory bring new opportunities and challenges for research on textual entailment. This paper clarifies the scope of textual entailment, gives a comprehensive survey on recognizing textual entailment and knowledge acquisition, according to previous held international evaluation workshops. The future research directions and new challenges are also elaborated under the current research situation.

In recent years, the authors' group has focused on the related researches with textual entailment, such as recognizing textual entailment, alignment for textual entailment, knowledge acquisition, information retrieval and question answering.

This work is supported by the National Natural Science Foundation of China (Nos. 61472105, 61472107), and the National High Technology Research and Development Program (863 Program) of China (No. 2015AA015407).