复述检测技术综述

李铂鑫[1,2], 李 鹏[1,2], 齐保元[1,2], 王 斌[1,2], 王丽宏[3]

[1]中国科学院信息工程研究所 北京 中国 100093
[2]中国科学院大学网络空间安全学院 北京 中国 100049
[3]国家计算机网络应急技术处理协调中心 北京 中国 100029

摘要 网络内容安全日益受到各界的关注。自然语言处理中用于判断两个文本语义是否相同的复述检测技术, 可以把语义相同表述形式不同的的看法、意见等聚成一类, 大幅提高舆情监控的效率; 亦可识别出经过改写的不良敏感信息, 有效提高不良敏感信息的召回率。本文旨在介绍当前复述检测技术领域的研究进展。首先介绍复述检测的概念、应用场景和研究现状。然后对复述检测方法进行分类, 本文从计算方式上将复述检测方法分为基于相似度的方法和基于特征的方法, 依次介绍每类方法的特点、优缺点, 并详述一些有代表性的方法, 重点介绍了基于深度学习的复述检测方法。最后详细分析了复述检测技术当前存在的问题, 并对未来的发展趋势进行了展望。

关键词 网络内容安全; 网络舆情监控; 自然语言处理; 复述检测; 深度学习; 神经网络
中图法分类号 TP391.1 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.09.07

A Survey on Paraphrase Identification Technology

LI Boxin[1,2], LI Peng[1,2], QI Baoyuan[1,2], WANG Bin[1,2], WANG Lihong[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[3]National Computer Network Emergency Response Technical Team Coordination Center of China, Beijing 100029, China

Abstract Network content security has received increasing attention from all walks of life. Paraphrase identification technology, commonly used to judge whether two text capture the same meaning in the field of natural language processing, can come in handy. This technology can aggregate the same views and opinions into the same category, greatly improving the efficiency of network public opinion monitoring. Also, it can identify the rewritten sensitive information and effectively improve the recall rate of bad sensitive information. This paper focuses on the research progress in the field of paraphrase identification. Firstly, we introduce the concept, application scenarios and research status of paraphrase identification. Secondly, we classify paraphrase identification methods into two categories: similarity-based methods and feature-based methods. Then we introduce the characteristics, advantages and disadvantages of each type in turn, and detail some representative ones. Among them, deep learning methods are highly focused. Finally, it is the detailed analysis of current problems and prospect of this field.

Key words web content security; public opinion monitoring; natural language processing; paraphrase identification; deep learning; neural network

1 引言

网络内容安全作为网络信息安全的一个重要组成部分,日益受到政府部门、工业界和学术界的关注。随着互联网特别是社交网络的高速发展, 自媒体和社交平台已经成为人们获取信息、发表意见、表达诉求、传播舆论的重要平台。与此同时, 网络违规现象愈加频繁: 大量不良敏感信息(如淫秽色情、反动言论、恐怖主义、恶意谣言等), 甚至一些机密信息会在互联网上进行传播, 严重影响了社会和政治稳定。因此, 网络舆情监控越来越重要, 不仅可以及时获取民意, 了解百姓对政策的反馈, 有效地提高政府的执政能力, 还能避免不良敏感信息等的传播, 维护社会稳定。

由于语言表达的灵活性, 同一个语义可能有多种不同的表述形式, 这给网络舆情监控带来了两个

通讯作者: 李鹏, 博士, 高级工程师, Email: lipeng@iie.ac.cn。
本课题得到国家重点研发计划课题(No. 2016YFB0801003); 中国科学院战略性先导科技专项(C 类)(No. XDC02040400)资助。
收稿日期: 2018-08-16; 修改日期: 2019-01-25; 定稿日期: 2020-07-31

挑战: 一是存在大量语义相同而表达方式不同的看法、意见等信息, 若进行有效处理, 会大幅度地增加相关人员的工作量; 二是人们出于躲避审查的目的, 往往会对待传播的不良敏感信息进行加工处理(如换种表达方式), 导致关键词匹配技术(一种广泛使用的内容分析与过滤技术)的失效。自然语言处理领域的复述检测技术, 可以判断两个文本的语义是否相同, 适于应对上述两个挑战。复述检测技术在网络信息安全中的作用, 主要体现在舆情聚类、不良敏感信息等的追踪方面, 可以匹配意思一样但表述不相同的信息, 有助于减少相关人员的工作量, 提高不良敏感信息等的召回率, 从而提升网络舆情监控的质量和效率。因此, 复述检测技术的研究对于网络信息安全是十分重要的。

本文首先介绍复述的概念。复述是一种常见的语言现象, 指的是用不同的词语来表达相同的语义。大部分语言学文献中所给出的复述的定义, 并不是严格意义上的语义相同, 而是一种宽泛的定义, 如“概念上的近似等价”[1-2]。 Bhagat 和 Hovy[3]把这种宽泛定义下的复述称为“准复述”(quasi-paraphrase)。本文所研究的复述也属于“准复述”。例如下面两句话, 虽然第一句中的“Authorities said”在第二句中没有相对应的部分, 但本文仍把它们视为复述句。

(1) Authorities said a young man injured Richard Miller.

(2) Richard Miller was hurt by a young man.

作为自然语言处理的一个重要领域, 复述研究不仅包括复述检测, 还包含复述语料库的构建和复述生成。其中, 复述检测指的是判断两个文本之间是否具有复述关系。复述语料库的构建指的是通过复述检测技术或简单的启发式规则, 生成包含大量复述文本对的数据集。复述生成指的是给定一个文本, 生成和它具有复述关系的新的文本。这三个任务之间相辅相成: 复述语料库的构建, 提供了数据集, 可以促进复述生成和复述检测技术的发展。另一方面, 复述检测技术的进步, 又可以反过来辅助复述语料库的构建, 提高复述生成的质量。

本文的主旨是对复述检测技术进行综述。由于复述包含多个粒度, 例如 Bazilay 等[4]将复述分为词汇级、短语级和句子级三类, 因此, 复述检测也对应多个级别, 不过, 本文只讨论句子级别的复述检测。另外, 同复述检测密切相关的任务是文本蕴含识别[5]。文本蕴含定义了文本 T(Text)和假设 H(Hypothesis)之间的二元关系, 具体而言, 如果能从文本 T 推理出假设 H, 就认为 T 蕴含 H。例如, 设文本 T 为“Mike bought some apples.”, 假设 H 为“Mike purchased some fruite.”, 则 T 蕴含 H。上述例子显示, 蕴含描述的是一种单向的关系, 因为无法从 H 推理出 T。而复述描述的是一种双向的关系, 通常被视作双向蕴含。例如, 句子 A 和 B 互为复述句, 则要求 A 蕴含 B, 同时 B 又蕴含 A。鉴于复述与蕴含之间存在上述联系, 便有研究人员把蕴含识别的相关方法, 推广到复述检测任务中[6]。

除了在网络信息安全领域的重要应用外, 复述检测技术在自然语言处理领域也有广泛的应用。例如, 在机器翻译中, 复述检测可以用于翻译结果的自动评价[7-9]。在自动文摘领域, 复述检测的作用体现在两个方面: 一是用于机器生成文摘的自动评价[10]; 二是在多文档自动文摘中, 用于冗余句子的去重。同单文档摘要相比, 多文档摘要在所选取的摘要句中, 更有可能出现语义相同的冗余句子的现象。因此, 为了提高摘要的质量, 冗余句子的去重是自动文摘, 尤其是多文档自动文摘任务中的一项必不可少的工作。而复述检测技术可以为冗余句子的发现提供线索。在问答系统中, 复述检测常用于答案句的抽取阶段, 可以匹配那些字面表述不同但语义相同的答案句, 以便找到更多匹配的答案句, 有助于提升答案句的召回率[11]。此外, 复述检测还可以自动发现文章的抄袭和剽窃现象[12]。

本文是关于目前复述检测技术研究进展的综述。文章的结构如下: 第二节介绍了复述检测的研究现状; 第三节为复述检测方法分类; 第四节介绍了基于相似度的复述检测方法; 第五节详述了基于特征的复述检测方法; 第六节介绍了常用的复述检测数据集; 第七节是存在的问题和展望; 最后是总结。

2 复述检测的研究现状

自 2004 年微软 Dolan 等人[13]发布了复述检测语料库 MSRP(MicroSoft Research Paraphrase Corpus)以来, 复述检测的研究方兴未艾, 逐渐成为研究热点。在这段时期内, 复述检测技术大致经历了三个发展阶段。

在第一个阶段, 研究人员从相似度的角度探索了句子间的复述关系[6, 20-24]。Fernand 等人[20]提出了一种矩阵相似度计算方法。Islam 等人[21]在计算句子相似度时, 统筹考虑了词与词之间的词汇相似度和语义相似度。Mihalcea 等人[22]广泛探索了 2 种基于语料库和 6 种基于知识的词相似度计算方式, 在此基础上提出了一种基于词相似度的短文本语义相似度计算方法。上述方法都着眼于词汇相似度, Milajevs 等人[23]把句子表示为向量, 以句子向量的余弦

值作为句子的相似度, 他们研究了不同的词向量表示, 并尝试了多种由词向量生成句子向量的组合方式。Guo 和 Diab[24]针对话题模型在计算句子语义相似度时的稀疏性问题, 提出了 WTMF(Weighted Textural Matrix Factorization)方法。Rus 等人[6]借助文本蕴含技术研究复述检测问题。

在第二个阶段, 研究人员开始利用机器学习技术研究复述检测问题。他们把复述检测当作一个分类任务, 通过人工构造特征, 在训练集上学出一个模型, 用于判断两个句子的复述关系。在该阶段中, 特征的构造是重点, 研究人员设计出了多种多样的特征。例如, 日本 ATR 研究所的 Finch 等人[26]和普林斯顿大学的 Madnani 等人[25]认为机器翻译结果的自动评价任务和复述检测任务很接近, 因此, 他们把机器翻译中的常见评价指标作为特征。Lintean 和 Rus[40]利用词汇特征(一个词或一个 bigram)构建了一个不相似核, 用于复述检测。Kozarev 和 Montoyo[27]使用了两类基于词汇的特征: 词重叠率特征和词相似度特征。Qiu 等人[28]利用句法特征研究了两个句子的不相同部分对复述关系的影响。鉴于复述检测是一个复杂的任务, 许多研究人员都混合使用了多种类型的特征。例如, Filice 和 Moschitti[29]在研究仅出现在一个句子中的文本片段是否会影响两个句子的复述关系时, 综合使用了语义特征(词向量)和句法特征。Liang 等人[30]研究了如何将局部特征和结构特征组合起来, 以便更好地学习文本的语义相似度。Ji 和 Eisenstein[31]综合使用了潜在语义特征、词汇特征和句法特征。Das 和 Smith[32]把句法特征和语义特征融合在一起, 提出了一种基于近似同步依赖语法的生成模型。Wan 等人[33]对复述检测相关的特征进行了分类, 划分为词汇特征、基于依赖的特征和句子长度特征, 重点探索了基于依赖的特征的优缺点。Yin 和 Schütze[34]把词嵌入思想推广到短语级别, 同时使用了语义特征和词汇特征。

在第三个阶段, 研究者把目光投向了深度学习。最早把深度学习技术引入复述检测领域的是斯坦福大学的 Socher 等人[18], 他们提出了一种基于递归自编码器的复述检测方法, 自动地学到了句子中短语和片段的表示。英国牛津大学的 Cheng 和 Kartsaklis[16]利用递归神经网络实现了动态词义消歧, 自动地学到了每个词的最佳语义向量。德国慕尼黑大学的 Yin 和 Schütze[15]提出了一种基于多粒度、交互式的 Bi-CNN-MI 模型, 在 4 种层面上提取特征。同样是使用了 CNN 结构, He 等人[14]提出的 Multi-Perspective CNN 模型, 不仅从粒度(卷积窗口的大小)出发, 还从卷积的方向、池化的类型、相似度计算策略多个角度, 抽取句子不同层面的特征。Yin 等人[17]第一次把基于注意力机制的 CNN 模型应用到自然语言处理领域。Kiros 等人[19]把词向量模型中的思想推广到句子级别, 根据书本中句子的连贯性, 通过目标句子重构上下文句子, 从而学到一种无监督的、通用的、分布式的句子向量表示。

国外的一些院校和研究机构较早、较广泛地对复述检测进行了研究。例如, 美国的斯坦福大学、普林斯顿大学、卡内基·梅隆大学、哥伦比亚大学、西北大学、马里兰大学、佐治亚理工学院、北德克萨斯大学、孟菲斯大学、德国慕尼黑大学、意大利罗马大学、加拿大多伦多大学、英国牛津大学、伦敦大学玛丽皇后学院、爱丁堡大学、新加坡国立大学、澳大利亚麦考瑞大学等高校, 以及 IBM 沃森研究中心、日本 ATR 研究院等研究机构都有相关研究人员对复述检测问题进行了研究。

表 1　复述检测方法分类

Table 1　Classification of paraphrase identification methods

基于相似度的方法		基于词汇相似度的方法[20-22]
		基于句子向量余弦相似度的方法[23-24]
基于特征的方法	基于人工构造特征的方法(传统机器学习方法)	基于词汇特征的方法[6,26-27,40]
		基于句法特征的方法[28]
		基于混合特征的方法[29-33]
	基于自动学习特征的方法(深度学习方法)	递归自编码模型: RAE[18]
		递归模型: SAMS-RecNN[16]
		卷积模型: Multi-Perspective CNN[14], Bi-CNN-MI[15], ABCNN[17]
		序列到序列模型: Skip-thought[19]

国内的高校和研究机构在复述方面有所探索。例如, 中国科学院自动化研究所的宗成庆探索了复述在中文口语中的应用[73-74]。哈尔滨工业大学的李维刚、赵世奇等人在复述实例和模板的抽取、复述

资源获取、复述生成以及复述应用等方面进行了许多尝试[75-81]。

总之, 国外对复述检测的研究集中于英文复述, 目前的主流方法是深度学习模型, 然而在常用的英文复述检测数据集 MSRP 上, 最好的成绩是传统的机器学习方法保持的[31], 深度学习模型仍有可提升的空间。国内在复述检测领域的研究还有待发力, 尤其是在中文复述检测方面, 根本原因在于缺乏一个公认的、高质量的中文复述检测数据集。

3 复述检测方法分类

随着自然语言处理技术的蓬勃发展, 深度学习技术的不断突破, 各种复述检测方法不断涌现。目前, MSRP 是复述检测领域公认的数据集, 大部分复述检测方法都是在该数据集上进行评测的, 同时也为了公平比较, 文中讨论的方法都基于该数据集。

复述检测的方法, 根据分类标准的不同, 可以划分为不同的类别。例如, 根据学习方式的不同, 分为有监督的方法和无监督的方法。根据是否使用外部资源, 可以分为使用外部资源的方法(我们认为预训练的词向量也属于外部资源)和不使用外部资源的方法。本文从计算方式上, 把复述检测方法分为两类: 基于相似度的方法和基于特征的方法。

基于相似度的方法, 可以根据相似度的计算方式的不同, 分为两类: 一是基于词汇相似度的; 二是基于句子向量余弦相似度的。

基于特征的方法也可以分为两类: 基于人工构造特征的方法和基于自动学习特征的方法, 分别对应传统机器学习方法和深度学习方法。其中, 基于人工构造特征的方法, 根据所使用的特征的类型, 又可分为基于词汇特征的方法、基于句法特征的方法和基于混合特征的方法。而基于自动学习特征的方法, 根据模型的类别, 可分为递归自编码模型(Recursive AutoEncoder: RAE)[18]、递归模型[16]、卷积模型(CNN)[14-15,17]和序列到序列模型[40]。表 1 显示的是本文中复述检测方法的分类体系。

4 基于相似度的方法

基于相似度的方法通过比较两个句子的相似度是否高于某个阈值, 来判断它们之间是否具有复述关系。其中, 阈值作为超参数, 可以人为给出, 一般设置为 0.5; 也可以在训练集或验证集上调试得到。句子相似度的计算方法可分为二类: 一是基于词汇相似度[20-22], 由词的相似度得到句子的相似度; 二是基于句子向量余弦相似度[23-24], 把两个句子向量的余弦值作为句子的相似度。

4.1 基于词汇相似度的方法

第一类方法通过探索不同的词相似度的计算方式, 基于“句子是由不同的词构成的”这一事实, 将句子的相似度看作词相似度的函数。常见的计算词相似度的方法包括: 基于词重叠率的方法, 基于知识的方法[48-53]和基于语料库的方法[54-56]。基于词重叠率的方法无法解决同义词问题。针对这一问题, 基于知识的方法和基于语料库的方法分别利用 WordNet[57]的层级结构和词共现原理, 可以计算出词与词之间的语义相似度, 有效地解决了同义词识别问题。有代表性的基于知识的方法包括: LCh[48]、Lesk[49]、WuP[50]、Res[51]、Lin[52]、JnC[53]等。而基于语料库的方法的代表有: PMI-IR[54]、LSA[55]、SOC-PMI[56]等。

文献[22]针对基于词重叠率的方法难以计算文本间的语义相似度这一问题, 例如“I own a dog. / I have an animal.”, 提出了一种基于词相似度的短文本语义相似度的计算方法。在计算词相似度时, 他们尝试了 2 种基于语料库的方法和 6 种基于知识的方法。实验结果表明, 文中提出的文本语义相似度方法, 好于简单的基于向量相似度方法。

文献[20]从信息抽取方法中得到灵感, 提出了一种矩阵相似度方法, 特点是在计算句子 A 与 B 的相似度时, 考虑了句子 A 中每个词与句子 B 中所有词的相似度, 而不同与其他方法, 只考虑最大的相似度。他们的方法在计算词相似度时, 用到了多种基于知识的方法。

文献[21]在计算句子相似度时, 综合了词与词之间的词汇相似度和语义相似度。其中, 词汇相似度采用归一化的最长公共子序列(NLCS: Normalized Longest Common Subsequence)来计算, 语义相似度基于语料库的方法 SOC-PMI[56]得出。

4.2 基于句子向量余弦相似度的方法

这类方法的核心思想是把句子表示为向量形式, 把两个句子向量的余弦值作为句子的相似度。

文献[23]探索了不同的词向量表示, 并尝试了多种由词向量生成句子向量的组合方式。实验结果表明: (1) 在小规模数据集上, 基于神经网络的词向量的性能优于(或至少接近于)基于共现的词向量; 基于张量的组合方式, 并不总是好于简单的组合方式。(2) 在大规模数据集上, 基于神经网络的词向量的性能明显优于基于共现的词向量。

话题模型常用于文档级别的语义相似度计算, 将其推广到句子语义相似度计算时, 会出现严重的稀疏

性问题。Guo 和 Diab 针对这一问题, 提出了一种名为 WTMF(Weighted Textural Matrix Factorization)的句子相似度计算方法[24]。他们认为句子中没有出现的词可以辅助句子表示。他们的方法与 WMF(Weighted Textural Matrix Factorization)方法类似, 通过分解矩阵得到句子的潜在语义向量, 两个潜在语义向量的相似度即为两个句子的相似度, 区别在于, 他们的方法对句子中没出现的词也进行了建模, 只不过在目标函数中, 给这些不可见词也分配了一个非常小的权重。该方法的优点是对句子中不可见词也进行了建模, 有效地解决了潜在语义模型对句子建模时的稀疏性问题; 缺点是没有考虑同义词问题。

4.3 小结

综上所述, 基于相似度的方法的优点是计算简单、无需训练、可解释性强。不过, 它的缺点也很明显, 主要体现在下述 4 个方面:

(1)没有考虑句法结构, 割裂了句子内部词与词之间的关系。例如, 会把下述例子误判为复述句: “John was hurt by Tom./Tom was hurt by John.”。

(2)无法解决同义短语的识别问题, 比如短语“take sth. into consideration/look after”。目前的方法一般只处理同义词问题。

(3)无法正确表示每个词在句子中的重要程度。现有的做法或者认为每个词在句子中的权重一样, 或者借用全局的 IDF 权重。

(4)没有考虑两个句子中的不相同成分, 对高重叠率的负例, 容易产生误判。例如, “The technology-laced Nasdaq Composite Index .IXIC added 1.92 points, or 0.12 percent, at 1,647.94./The technology-laced Nasdaq Composite Index .IXIC dipped 0.08 of a point to 1,646.”。

5 基于特征的方法

基于特征的方法把复述检测任务当作一个分类任务, 通过构造特征(人工特征或自动学习特征), 在训练集上学出一个模型, 用于复述检测。在复述检测领域, 绝大多数的方法都是有监督的。这些方法大致可以分为两类, 一类是基于人工构造特征的方法, 另一类是基于自动学习特征的方法。

5.1 基于人工构造特征的方法

基于人工构造特征的方法, 根据特征的类型, 又可分为基于词汇特征的方法、基于句法特征的方法和基于混合特征的方法。下面分别进行讨论。

5.1.1 基于词汇特征的方法

这类方法仅从词汇角度出发, 构造特征。例如, 文献[25-26]认为机器翻译结果的自动评价任务和复述检测任务很接近, 因此, 可以把机器翻译的评价指标作为特征, 训练分类器, 用于复述检测任务。文献[26]首次把机器翻译中的常用评价指标, 如 BLEU[64]、NIST[65]、WER[66]、PER[67], 引入到复述检测任务中, 取得了不错的效果。随着机器翻译技术的蓬勃发展, 又出现了一些新的机器翻译指标, 如 TER[68]、TERp(TER-Plus)[69]、METEOR[70]、SEPIA[71]、BADGER[72]、MAXSIM[45], 文献[25]探索了这些新的机器翻译指标对复述检测任务的影响, 融入了新的机器翻译指标后, 复述检测性能有了较大幅度的提升, 正确率和 F1 值分别提升了 2.4%和 1.4%。

文献[40]想法独特, 基于这样一种假设: 如果两个句子对应的不相似向量很接近, 那么这两个句子应该具有相同的复述关系。基于该假设, 文中提出了一种新颖的复述检测方法: 对一个实例(包含两个句子)而言, 构建一个不相似向量, 该向量中的非 0 维表示的是该维度上对应的元素(一个词或一个 bigram)只出现在其中的一个句子中; 然后利用文中提出的核函数, 在训练集上学习 SVM 分类器, 用于复述检测。

文献[27]使用了两类基于词汇的特征: 词重叠率特征和词相似度特征, 目的是探索下述三个问题: 一是哪类信息更有用; 二是把不同种类的信息组合起来, 效果是否会更好; 三是联合多种基于机器学习的方法, 进行投票表决, 性能是否有提升。

基于词汇特征的方法, 优点是操作简便; 缺点是忽略了词的语义信息, 以及词与词之间的关系, 会在一定程度上影响复述检测的效果。

5.1.2 基于句法特征的方法

句子中词与词之间的关系可以通过句法结构来描述。文献[28]首次从两个句子的不相同部分出发, 研究两者的复述关系, 认为在判断两个句子是否为复述句时, 需要考虑下述两个方面的情况: 一是它们是否拥有足够多的相同信息块; 二是那些不相同的信息块, 是否会影响两者的复述关系。文章通过语义角色标注工具, 将每个句子划分成若干个(谓词, 论元 0, 论元 1)这样的语义块, 认为两个句子中相似度高于某一阈值的语义块具有对齐关系, 把没对齐的语义块输入到一个专门训练的分类器中, 判断它们是否会影响复述关系。

基于句法特征的方法充分考虑了词语间的关系, 有助于识别语序不同导致语义不同的例子, 如“John was hurt by Tom./Tom was hurt by John.”。这类方法的缺点是依赖句法分析工具, 速度慢、会受句法分析工

具性能的影响而导致误差的二次传递。

5.1.3 基于混合特征的方法

复述检测是一个复杂的任务, 很难基于某一类单独的特征, 就能取得理想的效果。因此, 大多数方法都采用多种类型的特征, 如词汇特征、句法特征、语义特征等, 这类方法统称为基于混合特征的方法。容易看出, 该类方法融合了多种特征的优势, 效果往往比较好。

句法特征可以表示句子中词与词之间的关系, 是一类非常重要的信息, 可以解决语序不同导致的语义不一致的问题。因此, 不少方法中都用到了这类特征[29-33]。

Filice 和 Moschitti[29]在研究复述的一类特殊现象(即仅出现在一个句子中的文本片段, 是否会影响两个句子的复述关系)时, 除了使用词向量语义特征来计算词语间的相似度外, 还利用基于句法特征树结构的树核来组合特征。他们的做法比较有趣, 专门构造了一个数据集, 在上面训练出一个分类器, 用于判断仅出现在一个句子中的文本片段对两个句子间的复述关系是否有影响。然后, 他们将该分类器作为一个组件嵌入到其他复述检测方法中, 有助于提高其他方法的效果。

Liang 等人[30]基于句法树结构, 为每个句子生成了一个属性关系图(Attributed Relational Graph), 可以在该图上编码丰富的属性信息, 如节点的词性、词干、词向量、以及节点与节点间的依存关系等信息。他们研究的目的是如何有效地将局部特征和结构特征组合起来, 以便更好地学习文本的语义相似度。而特征的组合方式依赖于句子对齐, 他们巧妙地将句子对齐问题转换为结构预测问题, 提出了两种方案予以解决: 一是将句子对齐当作一个特征选择的过程; 二是将句子对齐看作一个潜变量。他们的方法效果还不错。另外, 鉴于充分利用了句法结构特征, 他们的方法特别适合解决语序不同导致的语义不一致的问题, 比如, 对于“A man plays a guitar./A guita plays a man.”这样的实例, 计算出的相似度仅为 0.3, 从而轻易地判断出两者不具备复述关系。

Ji 和 Eisenstein[31]综合使用了潜在语义特征、词汇特征和句法特征, 并提出了一种专门针对复述检测任务的权重计算方式 TF-KLD。虽然他们的方法在 MSRP 数据集上取得了当前最好结果, 准确率为 80.4%, F1 值为 85.9%。但是, 他们的最好效果是基于直推式(transductive)的方式, 该方式通过矩阵分解获得潜在语义特征的过程中, 使用了测试集的数据。如果不使用测试集数据, 即改用他们的归纳式(inductive)方式, 则在 MSRP 上的准确率为 77.8%, F1 值为 84.3%, 这一结果来自 Yin 和 Schütze[34]的工作。

Das 和 Smith[32]认为, 互为复述的两个句子, 其句法结构应该尽可能地对齐。他们提出的基于近似同步依赖语法的生成模型, 融合了句法特征和语义特征。

Wan 等人[33]将影响复述检测的特征分为 3 大类: 词汇特征、基于依赖的特征、以及句子长度特征, 重点探索了基于依赖的特征的优势和不足, 这些基于依赖的特征源自依存句法分析器。他们的实验结果表明, bigram 特征和基于依赖的特征, 在编码信息的能力是相当的, 这一结果正印证了 Collins (1996) 发现的现象: 在英语中, 大约 70%的依赖实际上是 2 个邻接的词。

当然, 也有一些基于混合特征的方法没有使用句法特征, 而是综合使用了语义特征和词汇特征[34-35]。例如, Yin 和 Schütze 把词嵌入思想推广到短语级别, 使用类似的策略学习短语的向量表示, 然后采用线性加权策略, 把词向量和短语向量组合起来, 得到句子的向量表示, 作为语义特征, 外加 Madnani 等人提出的 8 个机器翻译评价指标作为词汇特征, 取得了很不错的效果[34]。另外, 针对 Ji 和 Eisenstein 提出的 TF-KLD 方法无法计算未出现在训练集中的词的权重的问题, 他们提出了一种改进的名为 TF-KLD-KNN 的权重计算方法, 对于那些未出现的词或短语, 其权重由 k 近邻的平均权重确定。

5.1.4 小结

根据所使用的特征的类型, 我们将基于人工构造特征的方法划分为 3 大类: 基于词汇特征的方法、基于句法特征的方法和基于混合特征的方法。基于词汇特征的方法操作简便, 但忽略了词的语义信息, 以及词与词之间的关系, 会在一定程度上影响复述检测的效果。基于句法特征的方法优点是充分利用了词语间的关系, 缺点是过度依赖句法分析工具, 速度慢、会受句法分析工具性能的影响而导致误差的二次传递。基于混合特征的方法综合了多种特征的优势, 效果往往比较好。

基于人工构造特征的方法在复述检测任务上取得了很不错的结果, 但是它最大的缺点是需要人工设计特征, 这往往会耗费大量的人力物力, 尤其是在一些专业领域, 还需要专家的参与。遗憾的是, 人工很难设计出完全揭示数据集规律的所有特征, 尤其是那些隐藏在大规模数据集中的、对目标任务非常有帮助的特征。而深度学习方法正好善于捕捉这

些隐含的、非常有用的特征。

5.2 基于自动学习特征的方法

基于自动学习特征的方法属于深度学习范畴, 深度学习技术在计算机视觉[58-59]、语音识别[60]、以及自然语言处理中的机器翻译[61-62]、问答系统[63]等领域取得了的巨大成功, 也有不少研究人员尝试把深度学习技术应用到复述检测任务中[14-19]。

在复述检测领域, 深度学习与传统机器学习在特征提取方面相比, 优势主要体现在: (1)深度学习可以执行端到端的训练, 无需依靠语言学知识和经验去显式建模特征, 节省人力成本; (2)深度学习巨大的表征能力, 使数据通过网络高效地编码, 加之分布式词向量表示, 深度学习模型自身的层次化、序列化等结构, 非常适合建模自然语言, 可以有效地捕获自然语言中各种依赖关系等特征; (3) 深度学习模型利用大规模数据和高效算力的优势, 可以有效地捕捉数据内部不易被人工显式建模的隐含特征。

在本文中, 我们根据模型的类别, 把这些深度学习方法分为递归自编码模型[18]、递归模型[16]、卷积模型[14-15,17]和序列到序列模型[19]。我们首先从是否需要预训练、是否使用人工特征、词向量维数、准确率和 F1 值等角度, 列出了这些模型在微软复述检测数据集 MSRP 上的实验表现, 如表 2 所示。下面对每种模型中有代表性的方法进行详细介绍。

表 2 基于自动学习特征的复述检测方法在 MSRP 数据集上的实验结果

Table 2 Experimental results of the paraphrase identification methods based on automatic learning features on the MSRP corpus

	模型名称	预训练	人工特征	词向量维数	准确率(%)	F1 值(%)
递归自编码模型	RAE[18]	有	有	100	76.8	83.6
递归模型	SAMS-RecNN[16]	有	有	300	78.6	85.3
	Multi-Perspective CNN[14]	无	无	525	78.6	84.7
卷积模型	Bi-CNN-MI[15]	有	无	100	78.1	84.4
	ABCNN[17]	无	有	300	78.9	84.8
序列到序列模型	Skip-thought[19]	有	有	2400	75.8	83.0

5.2.1 递归自编码模型

斯坦福大学的 Socher 等人[18]在 2011 年提出了一种基于递归自动编码器(Recursive AutoEncoder: RAE)的复述检测方法。他们的基本思想是基于“编码器-解码器”框架, 依托句法分析树, 反向重构中间节点(或者直至叶节点), 把原始节点和重构节点对应的向量表示之间的欧氏距离和作为重构误差, 也就是模型要优化的目标函数, 在大规模无标注数据集上进行训练。最后, 只保留编码器端的参数, 即可得到句子中短语、片段、以及句子本身的向量表示。由于这些向量来自相同的语义空间, 因此, 可以方便地计算它们之间的语义相似度。基于句法分析树的重构过程如图 1 所示。该方法的最大特点是可以学到句子中的短语和片段的表示, 在计算两个句子的相似度时, 可以充分利用单词、短语、片段之间的两两相似度, 与之前只能利用单词与单词之间的相似度的做法不同。在计算出两个句子的相似度矩阵后, 作者利用动态池化 (dynamic pooling)技术, 将其转换成定长的特征矩阵, 作为深度学习模型提取的特征, 然后送入 Softmax 分类器, 便可判断两个句子的复述关系。该方法的基本流程如图 2 所示。

其中, 动态池化技术的目标是把变长的相似度矩阵, 转换成定长的特征矩阵。例如, 长度分别为 n 和 m 的句子 S_1 和 S_2, 对应的相似度矩阵为 $\boldsymbol{S} \in \mathbb{R}^{n \times m}$, 通过动态池化技术, 将其转换成定长的矩阵 $\boldsymbol{S}_{pooled} \in \mathbb{R}^{n' \times m'}$, 其中 n', m' 为预先给定的超参数。该过程相当于对不同的 S, 采用动态变化的池化核, 做无重叠的池化操作。此时, 池化核的宽度与高度分别为: $w = \lfloor n / n' \rfloor$, $h = \lfloor m / m' \rfloor$。存在 2 种特殊情况(以行为例, 列的情况类似): (1)不能整除时, 令 $M = n \bmod n'$, 此时多余的 M 行将均匀地分配在最后的 M 个池化窗口, 每个池化窗口会有 $\lfloor n / n' \rfloor$ 行; (2)句子过短时, 如 $n < n'$, 会通过上采样的技术, 重复采样行, 直到 $n' >= n$ 为止。

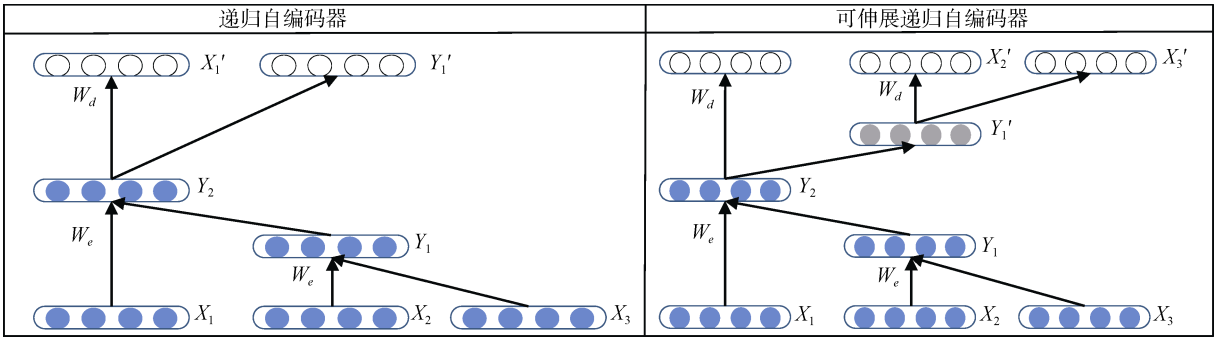

图 1　两种自动编码器. 左:标准的递归自编码器, 重构至子节点; 右: 可伸展递归自编码器, 重构至叶子节点[18]

Figure 1　Two types of autoencoders. Left: Standard recursive autoencoder, reconstructed to child nodes; right: Scalable recursive autoencoder, reconstructed to leaf nodes[18]

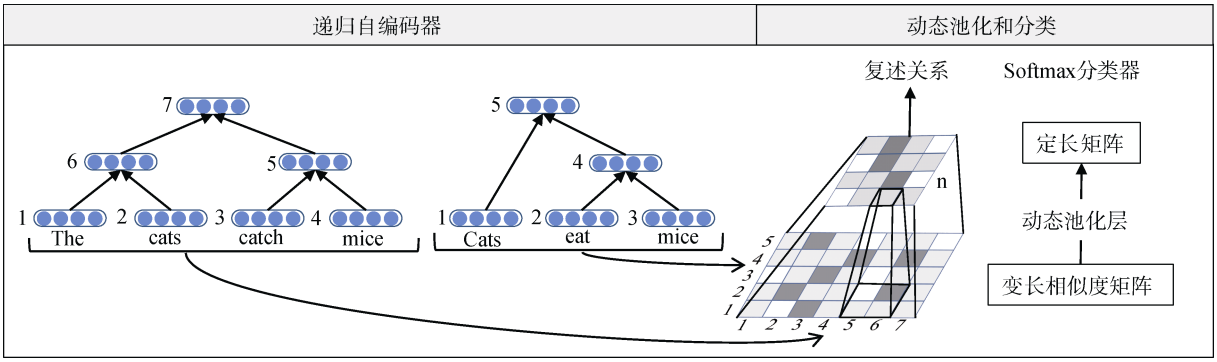

图 2　递归自编码模型[18]

Figure 2　The architecture of the recursive autoencoder[18]

实验结果表明(见表 3), 仅仅使用 RAE 和 DP (动态池化), 正确率只有 72.6%, 效果并不好。这是因为纯向量表示会丢失一些信息, 例如, 不同数字之间的向量表示非常相似, 而在 MSRP 数据集上, 数字间的很小差异, 就会导致两句话不是复述句。在结合了一些人工特征(如两个句子中所含的数字是否完全一样、两个句子的长度之差等)后, 模型正确率提高到了 76.8%, 取得了当时最好的结果。

表 3　递归自编码模型实验结果

Table 3　Experimental results of the recursive autoencoder

方法	正确率(%)	F1 值(%)
feats	73.2	
RAE+DP	72.6	
RAE+feats	74.2	
RAE+DP+feats	76.8	83.6

总的来说, 该模型的优点体现在两方面: (1)在计算句子间的语义相似度时, 考虑了不同粒度间的相似度(如词与词、词与词组、词组与词组); (2)通过使用动态池化技术, 成功地将维度不定的相似度矩阵, 转换为固定维度的特征矩阵, 既方便调用已有的机器学习模型, 又方便融合其他特征。而该模型的不足之处在于使用了句法分析工具, 容易引入噪声。

5.2.2　递归模型

递归模型(RecNN: Recursive Neural Network)与递归自编码模型(RAE)类似, 都需要借助外部的句法分析工具, 获取句子的句法分析树, 然后沿着句法树结构递归地组合词语, 得到句子的向量表示。不同点在于 RecNN 模型只做了 RAE 模型的编码过程, 得到句子的表示后, 直接进行预测输出, 用于下游任务的预测, 属于任务相关的模型; 而 RAE 还有解码阶段, 可在大规模无标注数据集上进行训练, 属于任务无关的模型。

英国牛津大学的 Cheng 和 Kartsaklis[16]在研究词汇歧义性对基于组合方式的句子表示的影响时, 探索了动态词义消歧与递归神经网络(RecNN: Recursive Neural Network)相结合的方式, 提出了一种句法感知多词义递归神经网络(Syntax-Aware Multi-Sense Recursive Neural Network: SAMS-RecNN)。他们给每个词分配多个数量相等的语义向量, 把词的语义向量和句子向量进行联合学习。至于最佳语义向量的

选择, 则是根据上下文动态进行的, 从而可以学到更好的句子表示。

5.2.3 卷积模型

德国慕尼黑大学 Yin 和 Schütze[15]认为, 对两个句子进行复述检测时, 需要在多个粒度上进行比较, 而卷积神经网络正适合提取不同粒度的信息。因此, 他们利用卷积结构, 提出了一种基于多粒度、交互式的Bi-CNN-MI模型, 分别在4种层面上提取特征: 单词级别, 短语级别, 长短语级别和句子级别。然后, 把 2 个句子不同层面上的特征两两求相似度, 得到 4 个相似度矩阵, 分别通过动态池化(dynamic pooling)操作, 转换成固定长度的特征矩阵, 拼起来作为逻辑斯蒂回归分类器的输入, 用于判断两者的复述关系。

他们的模型在复述检测任务上取得了相当不错的效果, 除了得益于不同粒度的信息丰富了句子的表示之外, 预训练在其中也起了重要作用。正如实验中所报告的, 没进行预训练时, 模型的准确率和 F1 值分别为 72.5%和 81.4%; 而预训练后, 模型的准确率和F1值则分别达到了为78.1%和84.4%, 效果提升非常明显。这主要是因为在 MSRP 这个小规模数据集上, 对复杂的模型进行训练时, 会存在两个问题: 一是数据稀疏, 二是可能导致过拟合。文中模型在大规模无标注数据集上的预训练过程, 有效克服了上述两个问题, 从而大幅提升了模型的效果。

同样是使用了 CNN 结构, He 等人[14]提出的 Multi-Perspective CNN 模型, 不仅从粒度(卷积窗口的大小)出发, 还从卷积的方向、池化的类型(min、max、avg)、相似度计算策略(cosin similarity、euclidean distance、absolute element-wise difference)多个角度, 抽取句子不同层面的特征, 拼接在一起, 构成最终的句子表示向量。其中, 卷积的方向分为 2 类: 匹配完整的词向量, 对应整体滤波器; 匹配词向量的每一维, 对应维度滤波器, 如图 3 所示。

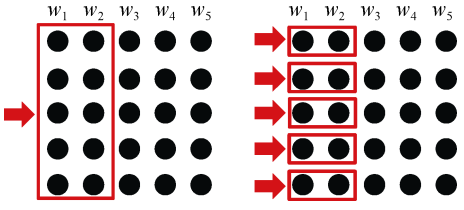

图 3 卷积方向示意图. 左: 整体滤波器, 匹配完整的词向量; 右: 维度滤波器, 匹配词向量的每一维[14]

Figure 3 The diagram of the convolutional direction. Left: the overall filter, matching the complete word vector; right: the dimensional filter, matching each dimension of the word vector[14]

除了从多个角度抽取句子不同层面的特征外, 在复述检测数据集上, 他们的方法还有 2 个特点: (1)使用多种类型的嵌入。除了使用常用的 GloVe 词嵌入[41]($Dim_g = 300$)外, 还用了 PARAGRAM 向量($Dim_k = 25$)和词性嵌入($Dim_p = 200$), 三者拼接起来, 作为词的最终表示: $Dim = Dim_g + Dim_k + Dim_p = 525$. 其中, PARAGRAM 向量是在大规模复述短语数据集 Paraphrase Database[42]上, 通过无监督方式训练出来的, 用于复述任务的词嵌入。词性嵌入是通过 word2vec 工具集[43],在新华机器翻译平行语料上训练出来的, 首先会利用斯坦福的词性标注器[44], 对语料中的英文句子标注词性。(2)采用了合页损失函数, 相比较对数损失函数而言, 合页损失函数更简单, 因为它只惩罚误分的实例:

$$loss(\theta, x, y_{gold}) = \sum_{y' \neq y_{gold}} \max(0, 1 + f_\theta(x, y') - f_\theta(x, y_{gold}))$$

其中, x 表示句子对: $x = (S_1, S_2)$, y_{gold} 是实例的实际标签, y_{gold} 是模型预测的标签, θ 是模型的参数, $f_\theta(x, y)$ 表示模型的输出。实验结果表明, Multi-Perspective CNN 模型在没进行预训练, 并且没使用人工特征的条件下, 效果仍好于上述 Bi-CNN-MI 模型。究其原因, 可能是因为 Multi-Perspective CNN 模型的输入融合了多种类型的嵌入, 使得表示更加丰富; 不过, 论文中并没有报告单独使用词嵌入时的实验结果。

Yin 等人[17]第一次把基于注意力机制的 CNN 模型应用到自然语言处理领域。他们提出了一种名为 ABCNN (Attention-Based Convolutional Neural Network) 的模型, 用于句子关系的建模。除了深度学习模型生成的特征外, 他们还额外使用了句子长度特征, 以及 15 个机器翻译特征。

5.2.4 序列到序列模型

Kiros 等人[19]从词向量模型 skip-gram[43]中得到灵感, 将其中的思想由词级别推广到句子级别, 根据书本中句子的连贯性, 通过目标句子重构上下文句子, 从而学习一种无监督的、通用的、分布式的句子向量表示。他们把这个模型称为“skip-thoughts”, 生成的句子向量称作“skip-thoughts vectors”, 模型如图 4 所示。

文中模型生成的句子向量在语义相似度、复述检测、文本分类等自然语言处理任务里进行了广泛的评价。和 RAE 模型类似, 基于无监督方式得到的

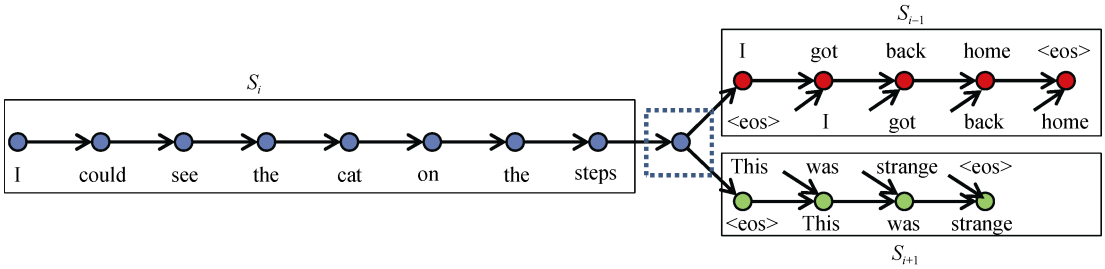

图 4　skip-thoughts 模型示意图。 三元组 (s_{i-1}, s_i, s_{i-1}) 表示本书中 3 个连续的句子, 模型利用“编码器-解码器”框架, 根据第 *i* 个句子, 分别预测第 *i–1* 个句子和第 *i+1* 个句子[19]

Figure 4　The overview of the skip-thoughts model. The triples represent 3 consecutive sentences in the book. The model uses the “encoder-decoder” framework to predict the (*i–1*)-*th* sentence and the (*i+1*)-*th* sentence respectively according to the *i-th* sentence[19]

句子向量所构造的特征, 表示能力有限, 在复述检测任务上效果并不好, 最佳模型的准确率也只有 73.0%, 在融合了 Socher 等人[18]设计的人工特征后, 准确率提高到了 75.8%。

在对序列进行建模时, skip-thoughts 模型没有使用 LSTM(Long Short Term Memory)[46], 而是使用计算更加简单的 GRU(Gated Recurrent Unit)[47]作为编码器和解码器网络结构的基本单元。GRU 是循环神经网络(RNN: Recurrent Neural Network)的一种变体, 和 LSTM 一样, 也是通过门的方式来控制读写。

在编码器端, GRU 的实现方式表示为:

$$r^t = \sigma(W_r x^t + U_r h^{t-1}),$$
$$z^t = \sigma(W_z x^t + U_z h^{t-1}),$$
$$\bar{h}^t = tanh(W x^t + U(r^t \odot h^{t-1})),$$
$$h^t = (1 - z^t) \odot h^{t-1} + z^t \odot \bar{h}^t$$

其中, z^t 为更新门, r_t 为重置门, $\bar{h}^t$ 表示的是 t 时刻待更新的状态, h^t 为 t 时刻的输出。

在解码器端, GRU 的实现方式与编码器端很相似, 区别在于额外增加了编码器的输出, 用公式表示为(以第 *i+1* 个句子的生成为例, 部分公式省去了下标):

$$r^t = \sigma(W_r^d x^t + U_r^d h^{t-1} + C_r h_i),$$
$$z^t = \sigma(W_z^d x^t + U_z^d h^{t-1}) + C_z h_i),$$
$$\bar{h}^t = tanh(W x^t + U^d (r^t \odot h^{t-1}) + C h_i),$$
$$h_{i+1}^t = (1 - z^t) \odot h^{t-1} + z^t \odot \bar{h}^t$$

skip-thoughts 模型的优点在于可以使用大量无标注数据进行训练, 生成的句子向量与具体任务无关, 可以应用在包括复述检测在内的多个自然语言处理任务中。它的缺点是模型较复杂, 导致训练时间过长, 尤其是当句子过长时, 这一缺陷更加明显。

5.2.5　小结

本节中, 我们根据模型的类别, 把基于自动学习特征的复述检测方法分为递归自编码模型、递归模型、卷积模型和序列到序列模型, 并介绍了每种模型中有代表性的方法。目前, 虽然深度学习技术在复述检测领域取得了很大的进展, 然而, 在常用的复述检测数据集 MSRP 上, 仅仅通过深度学习技术提取的特征, 效果仍没有好于传统的人工设计的特征。一个可能的解释是深度学习技术并没有充分提取适合于复述检测任务的特征, 这主要是因为 MRSP 数据集过小, 直接在它上面训练深度模型, 很容易出现过拟合现象; 而基于预训练的迁移学习, 又没能找到非常合适的迁移任务。因此, 大规模复述检测语料库的构建、合适的迁移任务的选取必将极大地促进深度学习技术在复述检测领域的发展。

6　复述检测常用语料

Dolan 等人从互联网上跨度为 18 个月的数以千计的新闻中, 通过一定的技术手段, 例如编辑距离、启发式规则(认为两个相似的新闻报道中, 第一句话很有可能互为复述句, 因为它们大都是文章的摘要), 选取了 5801 个句子对, 让两个人单独判断它们之间是否存在复述关系, 出现的分歧由第三个人负责解决。通过这种方式, 共有 3900 个句子对被判为复述句, 约占原始句子对的 67%。他们随后把这标注好的 5801 个句子对, 随机划分为两部分, 其中, 训练集占 70%, 共 4076 个句子对; 测试集占 30%, 共 1725 个句子对。上述过程构建的数据集, 称为 MSRP 数据集, 该数据集中的样例有两个显著特点: 一是正例(互为复述句)的词重叠率较低; 二是负例(不是复述句)的词重叠率较高。这使得朴素的复述检测方法, 如基于词相似度的方法等, 在该数据集

上表现欠佳, 从而促进了更深层次的、基于语义的复述检测方法的发展。

7 存在的问题和展望

(1) 缺乏大规模的复述检测语料库。目前公认的复述检测语料库是 Dolan 等在 2004 年构建的, 该数据集规模非常小, 只有 5801 个句子对。而当下如火如荼的深度学习技术需要大量的数据, 虽然在复述检测领域也出现了一些深度学习模型 , 但它们一般都需要预先在大规模无监督语料(或具有大规模数据的有监督任务)上进行预训练。总之, 数据集的短缺严重制约了复述检测技术的发展。

最近, 随着机器翻译性能的大幅提升, Wieting 和 Gimpel[12]利用神经机器翻译模型, 基于捷克文-英文的双语语料, 将其中的捷克文句子翻译成英文, 同原始英文句子一道, 构成复述句, 从而形成了千万级别的大规模复述语料。遗憾的是, 他们所构建的复述语料只有正例, 而缺乏负例, 并不适合复述检测任务。虽说可以利用随机负采样等方式来构建负例, 但通过这些方式得到的负例, 相似度低, 干扰性差, 很容易被一般的复述检测模型识别。随着文本生成模型的迅猛发展[37-38], 期望可以借助文本生成技术, 对原始句子进行编辑、改写, 以期自动生成大规模、高质量的负例, 从而推动复述检测技术取得更大进展。

(2) 很少从推理角度来研究复述检测。有些类型的复述需要经过推理才能识别出来, 例如下面两句话“Doctors had planned to deliver him two weeks early , on or around November 14.”和“A Caesarean had originally been planned in mid-November , two weeks early.”。这种类型的复述需要借助于规则和外部知识才可以推理出来。而目前的复述检测方法大多是基于句子语义相似度的, 很难识别出上述复述现象。

(3) 目前的复述检测方法大多是基于句子语义相似度的, 很少专注于复述现象本身。正如我们在前文中所言, 复述与语义相似度不是等价的, 语义相似度高的句子并不一定具有复述关系。因此, 仅仅通过两个句子的语义相似度来判断两者的复述关系是不够的, 还需要结合复述现象的特点进行更深入的比较。另外, 当前评价复述检测方法的性能优劣时, 所采用的指标是测试集上的 F1 值和正确率, 即整体性能, 而复述是包含多种类别的[40], 识别不同类别复述现象的难易程度是不一样的。因此, 如果能构建出既标注两个句子的复述关系, 又包含它们所对应的复述类型的复述检测数据集, 无疑会加快复述检测技术的发展。

8 总结

复述检测问题是自然语言处理领域的一类重要问题。本文从复述检测的应用场景、常见特征、主流方法、常用语料库等多方面对复述检测技术进行了综述。文章的重点是关于复述检测方法的综述, 把复述检测方法划分为两个大类, 多个子类, 通过对每个类别中最具代表性的方法的介绍, 分析出每类方法的特点以及优缺点, 希望对今后的研究者有所启迪。

复述检测作为自然语言处理中的一项基础任务, 迄今为止, 已经进行了广泛的研究, 但仍然存在大规模复述语料库的缺乏、基于推理角度的复述检测的研究的不足、以及很少专注复述现象本身的研究等问题。另外, 目前有关复述检测的研究大多是针对英文的, 针对中文的复述检测很少见到, 根本原因在于没有公认的中文复述检测语料库。因此, 为了推进中文复述检测的研究, 一个规模较大、被研究人员广泛认可的中文复述检测语料库的构建是必不可少的。前文提到的基于机器翻译系统获取英文复述语料的方法, 也许可以借鉴过来, 用于构建中文复述语料库。相信随着句法解析、语义角色标注等底层的自然语言处理技术的发展, 以及深度学习等机器学习技术的进步, 复述检测技术也必将迎来更大的发展。同时, 复述检测技术的发展, 也必将推动机器翻译、自动文摘、问答系统等其他更高一层的自然语言处理任务的进步。

参考文献

[1] de Beaugrande R, Dressler W U. Introduction to Text Linguistics[M]. Routledge, 1981.

[2] M.A.K. Halliday. An introduction to functional grammar[M]. London: Edward Arnold, 1994:110.

[3] Bhagat R, Hovy E. What is a Paraphrase?[J]. *Computational Linguistics*, 2013, 39(3): 463-472.

[4] R. Barzilay, K. R. Mckeown. Extracting paraphrases from a parallel corpus[C]. *Association for Computational Linguistics*, *ACL*, 2001: 50-57.

[5] Dagan I, Glickman O. Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability[J]. *In: Learning Methods for Text Understanding and Mining*, 2004: 26-29.

[6] V. Rus, P. M. Mccarthy, M. C. Lintean, et al. Paraphrase Identification with Lexico-Syntactic Graph Subsumption[C]. *International*

Florida Artificial Intelligence Research Society Conference, 2008: 201-206.

[7] Kauchak D, Barzilay R. Paraphrasing for Automatic Evaluation[C]. *the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006: 455-462.

[8] Zhou L, Lin C Y, Hovy E. Re-evaluating Machine Translation Results with Paraphrase Support[C]. *the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006: 77-84.

[9] Y. Lepage, E. Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation[C]. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)* , 2005: 57−64.

[10] L. Zhou, C. Y. Lin, D. S. Munteanu, et al. Paraeval: Using paraphrases to evaluate summaries automatically[C]. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, *ACL*, 2006: 447−454.

[11] Z. Wang, H. Mi, A. Ittycheriah. Sentence similarity learning by lexical decomposition and composition[C]. *International Conference on Computational Linguistics*, 2016: 1340−1349.

[12] Sánchez-Vega F, Villatoro-Tello E, Montes-Y-gómez M, et al. Paraphrase Plagiarism Identification with Character-level Features[J]. *Pattern Analysis and Applications*, 2019, 22(2): 669-681.

[13] Dolan B, Quirk C, Brockett C. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources[C]. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004: 350-356.

[14] He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks[C]. *Proceedings of the 2015 conference on empirical methods in natural language processing,* 2015: 1576-1586.

[15] W. Yin, H. Schütze. Convolutional neural network for paraphrase identification[C]. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(HLT-NAACL2015)*, 2015: 901-911.

[16] J. Cheng, D. Kartsaklis. Syntax-aware multi-sense word embeddings for deep compositional models of meaning[C]. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing(EMNLP2015)*, 2015: 1531-1542.

[17] W. Yin, H. Schütze, B. Xiang, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs[J]. *Transactions of the Association for Computational Linguistics*, 2016, 4: 259-272.

[18] R. Socher, E. H. Huang, J. Pennin J, et al. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection[C]. *Advances in neural information processing systems (NIPS2011)* , 2011: 801-809.

[19] R. Kiro, Y. Zhu, R. R. Salakhutdinov, et al. Skip-thought vectors[C]. *Advances in neural information processing systems(NIPS2015)*, 2015: 3294-3302.

[20] S. Fernand, M. Stevenson. A semantic similarity approach to paraphrase detection[C]. *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 2008: 45-52.

[21] A. Islam, D. Inkpen Semantic similarity of short texts[C]. *Recent Advances in Natural Language Processing*, 2009: 227-236.

[22] R. Mihalcea, C. Corley, C. Strapparava. Corpus-based and Knowledge-based Measures of Text Semantic Similarity[C]. *National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference(AAAI2006)*, 2006: 775-780.

[23] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, et al. Evaluating neural word representations in tensor-based compositional settings[C]. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP2014)*, 2014: 708-719.

[24] W. Guo, M. Diab. Modeling sentences in the latent space[C]. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics(ACL2012)*, 2012: 864-872.

[25] N. Madnan, J. Tetreaul, M. Chodorow. Re-examining machine translation metrics for paraphrase identification[C]. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(HLT-NAACL2012)*, 2012: 182-190.

[26] A. Finch, Y. S. Hwang, E. Sumita. Using machine translation evaluation techniques to determine sentence-level semantic equivalence[C]. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[27] Kozareva Z, Montoyo A. Paraphrase Identification on the Basis of Supervised Machine Learning Techniques[J]. *Advances in Natural Language Processing*, 2006: 524-533. DOI:10.1007/11816508_52.

[28] L. Qiu, M. Y. Kan, and T. S. Chua. Paraphrase recognition via dissimilarity significance classification[C]. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing(EMNLP2006)*, 2006: 18-26.

[29] S. Filice, A. Moschitti. Learning to Recognize Ancillary Information for Automatic Paraphrase Identification[C]. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL2016)*, 2016: 1109-1114.

[30] C. Liang, P. K. Paritosh, V. Rajendran, et al. Learning Paraphrase Identification with Structural Alignment[C]. *International Joint Conference on Artificial Intelligence (IJCAI2016)* , 2016:

2859-2865.

[31] Y. Ji, and J. Eisenstein. Discriminative improvements to distributional sentence similarity[C]. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing(EMNLP2013)*, 2013: 891-896.

[32] D. Das, N. A. Smith. Paraphrase identification as probabilistic quasi-synchronous recognition[C]. *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*, 2009: 468-476.

[33] Wan S, Dras M, Dale R, et al. Using Dependency-Based Features to Take the "Para-farce" out of Paraphrase[J]. *Proceedings of the Australasian Language Technology Workshop (ALTW 2006)*, 2006(2005): 131-138.

[34] W. Yin, and H. Schütze. Discriminative Phrase Embedding for Paraphrase Identification[C]. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(HLT -NAACL2015)*, 2015: 1368-1373.

[35] Ul-Qayyum Z, Altaf W. Paraphrase Identification Using Semantic Heuristic Features[J]. *Research Journal of Applied Sciences, Engineering and Technology*, 2012, 4(22): 4894-4904.

[36] W. Blacoe, M. Lapata. A comparison of vector-based representations for semantic composition[C]. *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning(EMNLP2012& CONLL2012)*, 2012: 546-556.

[37] Guu K, Hashimoto T B, Oren Y, et al. Generating Sentences by Editing Prototypes[EB/OL]. 2017: arXiv:1709.08878[cs.CL]. https://arxiv.org/abs/1709.08878.

[38] Gupta A, Agarwal A, Singh P, et al. A Deep Generative Framework for Paraphrase Generation[EB/OL]. 2017: arXiv:1709.05074. https://arxiv.org/abs/1709.05074.

[39] S. Q. Zhao, T. Liu, S. Li. Research on Paraphrasing Technology[J]. *Journal of Software*. 2009, 20(8): 2124-2137.
(赵世奇, 刘挺, 李生. 复述技术研究[J]. *软件学报*, 2009, 20(8): 2124-2137.)

[40] M. C. Lintean, V. Rus. Dissimilarity Kernels for Paraphrase Identification[C]. *Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 2011.

[41] J. Pennington, R. Socher R, C. Manning. Glove: Global vectors for word representation[C]. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP2014)*, 2014: 1532-1543.

[42] J. Wieting, M. Bansal, K. Gimpel, et al. From paraphrase database to compositional paraphrase model and back[J]. *Computer Science*, 2015: 98-104.

[43] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[EB/OL]. 2013: arXiv:1301.3781. https://arxiv.org/abs/1301.3781.

[44] C. Manning, M. Surdeanu, J. Bauer, et al. The Stanford CoreNLP natural language processing toolkit[C]. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014: 55-60.

[45] Y. S. Chan, H. T. Ng. MAXSIM: A maximum similarity metric for machine translation evaluation[C]. *Proceedings of the Meeting of the Association for Computational Linguistics*, 2008: 55-62.

[46] S. Hochreiter, and J. Schmidhuber. Long short-term memory[J]. *Neural computation,* 1997, 9(8): 1735-1780.

[47] Chung J, Gulcehre C, Cho K, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[EB/OL]. 2014: arXiv:1412.3555. https://arxiv.org/abs/1412.3555.

[48] C. Leacock, M. Chodorow. Combining local context and WordNet similarity for word sense identification[J]. *WordNet: An electronic lexical database*, 1998, 49(2).

[49] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone[C]. *Proceedings of the 5th annual international conference on Systems documentation*, 1986: 24-26.

[50] Wu Z B, Palmer M. Verbs Semantics and Lexical Selection[C]. *the 32nd annual meeting on Association for Computational Linguistics*, 1994: 133-138.

[51] Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy[EB/OL]. 1995: arXiv:cmp-lg/9511007. https://arxiv.org/abs/cmp-lg/9511007.

[52] D. Lin. An information-theoretic definition of similarity[C]. *International Conference on Machine Learning(ICML1998)*, 1998: 296-304.

[53] Jiang J J, Conrath D W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[EB/OL]. 1997: arXiv:cmp-lg/9709008. https://arxiv.org/abs/cmp-lg/9709008.

[54] P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL[C]. *European Conference on Machine Learning*, 2001: 491-502.

[55] Landauer T K, Foltz P W, Laham D. An Introduction to Latent Semantic Analysis[J]. *Discourse Processes*, 1998, 25(2/3): 259-284.

[56] A. Islam, D. Inkpen. Second order co-occurrence PMI for determining the semantic similarity of words[C]. *the International Conference on Language Resources and Evaluation*, 2006: 1033-1038.

[57] Miller G A. WordNet: A Lexical Database for English[J]. *Communications of the ACM*, 1995, 38(11): 39-41.

[58] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition[C]. *Proceedings of the IEEE conference on computer*

vision and pattern recognition, 2016: 770-778.

[59] J. Liu, G. Wang, P. Hu, et al. Global context-aware attention lstm networks for 3d action recognition[C]. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017)*, 2017: 43.

[60] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 82-97.

[61] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[EB/OL]. 2014: arXiv:1409.0473[cs.CL]. https://arxiv.org/abs/1409.0473.

[62] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need[C]. *Advances in Neural Information Processing Systems(NIPS2017)*, 2017: 5998-6008.

[63] M. Feng, B. Xiang, M. R. Glass, et al. Applying deep learning to answer selection: A study and an open task[C]. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU2015)*, 2015: 813-820.

[64] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]. *the 40th Annual Meeting on Association for Computational Linguistics*, 2001: 311-318.

[65] Doddington G. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics[C]. *the second international conference on Human Language Technology Research*, 2002: 138-145.

[66] Su K Y, Wu M W, Chang J S. A New Quantitative Quality Measure for Machine Translation Systems[C]. *the 14th conference on Computational linguistics*, 1992: 433-439.

[67] C. Tillmann, S. Vogel, H. Ney, et al. Accelerated DP based search for statistical translation[C]. *European Conf on Speech Communication & Technology*, 1997: 2667-2670.

[68] M. Snover, B. Dorr, R. Schwartz, et al. A study of translation edit rate with targeted human annotation[C]. *Proceedings of association for machine translation in the Americas*, 2006: 223-231.

[69] Snover M G, Madnani N, Dorr B, et al. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate[J]. *Machine Translation*, 2009, 23(2/3): 117-127.

[70] M. Denkowski, A. Lavie. Extending the METEOR Machine Translation Metric to the Phrase Level for Improved Correlation with Human Post-Editing Judgments[C]. *the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(HLT-NAACL2010)*, 2010: 250-253.

[71] N. Habash, A. Elkholy. SEPIA: surface span extension to syntactic dependency precision-based MT evaluation[C]. *Proceedings of the NIST metrics for machine translation workshop at the association for machine translation in the Americas conference (AMTA-2008)*, 2008:25-31.

[72] S. Parker. BADGER: A new machine translation metric[J]. *Metrics for Machine Translation Challenge*, 2008, 21-25.

[73] Zong C Q, Zhang Y J, Yamamoto K, et al. Approach to Spoken Chinese Paraphrasing Based on Feature Extraction[C]. *Nlprs*, 2001:551-556.

[74] C. Zong. Paraphrasing Chinese utterances in spoken language translation system[C]. *International Conference of Chinese Computing*, 2001: 395-401.

[75] S. Q. Zhao. Pivot approach for extracting paraphrase patterns from bilingual corpora[C]. *the 46th Annual Meeting of the Association for Computational Linguistics(ACL20108)*, 2008: 780-788.

[76] W. G. Li. Automated generalization of phrasal paraphrases from the Web[C]. *the Third International Workshop on Paraphrasing (IWP2005)*, 2005: 49−56.

[77] T. Liu, W. G. Li, Y. Zhang, et al. A Survey on Paraphrasing Technology[J]. *Journal of Chinese Information Processing*, 2006, 20(4): 25-32.
(刘挺, 李维刚, 张宇, 等. 复述技术研究综述[J]. *中文信息学报*, 2006, 20(4): 25-32.)

[78] 李维刚. 中文复述实例与复述模板抽取技术研究[PD]. 哈尔滨: 哈尔滨工业大学, 2008.

[79] Lazaridou A, Pham N T, Baroni M. Combining Language and Vision with a Multimodal Skip-gram Model[C]. *the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015: 1021-1029.

[80] S. Q. Zhao. Learning question paraphrases for QA from Encarta logs[C]. *International Joint Conference on Artificial Intelligence (IJCAI2007)*, 2007: 1796-1800.

[81] S. Q. Zhao. Automatic acquisition of context-specific lexical paraphrases[C]. *International Joint Conference on Artificial Intelligence (IJCAI2007)*, 2007: 1789-1794.

李铂鑫 男, 于 2012 年在华中科技大学计算机科学与技术专业获得硕士学位。现在中国科学院信息工程研究所攻读博士学位。研究领域为自然语言处理。研究兴趣包括: 深度学习、信息检索。Email: liboxin@iie.ac.cn

李鹏 男, 于 2013 年在中国科学院计算所计算机科学与技术专业获得博士学位。现在中国科学院信息工程研究所任高级工程师。研究领域为信息检索。研究兴趣包括: 自然语言处理。 Email: lipeng@iie.ac.cn

齐保元 男, 于 2015 年在中科院计算所计算机软件与理论获得博士学位。现在中国科学院信息工程研究所任助理研究员。研究领域为机器学习与舆情计算。 Email: qibaoyuan@iie.ac.cn

王斌 男, 于 1999 年在中科院计算所获得博士学位。现在中国科学院信息工程研究所任研究员, 中国科学院大学岗位教授。研究方向为信息检索与文本挖掘, 主要研究信息检索的理论、模型、算法、关键技术及其在文本处理领域中的应用。Email: wangbin@iie.ac.cn

王丽宏 女, 1967 年生, 博士, 国家计算机网络应急技术处理协调中心副总工程师, 研究方向为信息检索和智能信息处理。Email: wlh@isc.org.cn