自然语言处理发展及应用综述

赵京胜* 宋梦雪 高 祥
ZHAO Jing-sheng SONG Meng-xue GAO Xiang

摘 要 自然语言处理旨在设计算法使计算机像人一样理解和处理自然语言，是互联网和大数据时代的必然。自然语言处理涉及许多领域，包括词汇、句法、语义和语用分析，文本分类、情感分析、自动摘要、机器翻译和社会计算等。随着通信和计算机相关技术的发展，自然语言处理的应用需求也越来越大。分析自然语言处理的相关背景、常用方法和应用领域，并对NLP的发展进行了展望。

关键词 自然语言处理；信息抽取；自动文摘

doi：10.3969/j.issn.1672-9528.2019.07.046

1 前言

人类的日常生活离不开语言，自然语言作为一种最直接和简单的表达工具无处不在，自然语言处理（Natural Language Processing，NLP）是将人类交流沟通所用的语言经过处理转化为机器所能理解的机器语言，是一种研究语言能力的模型和算法框架，是语言学和计算机科学的交叉学科。作为人工智能的一个重要分支，在数据处理领域也占有越来越重要的地位，如今被大多数人熟知和应用。自然语言处理主要分两个流程：自然语言理解（Natural language Understanding ，NLU）和自然语言生成（Natural language Generation，NLG）。NLU主要是理解文本的含义，具体到每个单词和结构都需要被理解；NLG与理解相反，分三个阶段，确定目标，通过评估情况和可用的交际资源来计划如何实现目标，并将计划形成为文本。

本文对自然语言处理的相关概念、发展历史和相关研究问题进行分析，特别是自然语言处理应用领域的知识体系，包括文本分类、自动文摘等领域，并对自然语言处理的发展进行了展望和预测。

2 自然语言处理的发展

自然语言处理是一门包含着计算机科学、人工智能以及语言学的交叉学科，这些学科既有区别又相互交叉。其发展历程可分为四个阶段：1956年以前的萌芽期，1957-1970年是快速发展期，1971-1993年是低谷发展期，1994年到如今是复苏融合期。

1936年A.M.Turing发明了“图灵机”，使纯数学的逻辑符号和实体世界之间建立了联系，为后来计算机的发展提供了理论基础。20世纪50年代提出的自动机理论以图灵机的计算模型为基础，被认为是现代计算机科学发展的基础[1]。后来Kleene又在这种模型之上提出了有限自动机和正则表达式。1956年，Chomsky提出了上下文无关语法，同年在人工智能诞生之后，自然语言处理迅速融入该领域之中。在快速发展期，上下文无关语法的提出使得该领域的研究分为了基于规则的符号派和基于概率的随机派[2]，促使了未来的很多年人们都在研究这两种方法到底哪种方法更有效。在低谷期，许多研究人员也在一直坚持并取得了一些成果，70年代的语音识别算法研制成功，隐马尔科夫模型（Hidden Markov Model，HMM）提出并得到了广泛应用[1]。繁荣期主要表现在三个方面：首先是概率方法的大规模应用；其次是计算机的速度和存储量的大幅度提高，促使该领域的物质基础得到了改善；最后是网络技术的发展带来的强大推动力。

3 自然语言处理的研究方法和内容

3.1 自然语言处理的研究方法

中文信息处理主要是对字、词、段落或篇章进行处理。主要方法分别是基于规则和基于统计的方法，前者是人工根据语言相关的规则对文本进行处理；后者则是通过大规模的数据库分析数据，从而实现对自然语言的处理。自然语言处理受数据影响较大，而数据的增长是大多数NLP应用（如机器翻译）性能提高的原因，所以拥有强大的数据支持才可以更好的对文本进行进一步的理解和分析，这使得如今很多NLP应用程序采用数据流分析方法[3]。

自然语言的处理流程大致可分为五步：第一步获取预料。第二步对语料进行预处理，其中包括语料清理、分词、

* 青岛理工大学 山东青岛 266520

词性标注和去停用词等步骤。第三步特征化，也就是向量化，主要把分词后的字和词表示成计算机可计算的类型（向量），这样有助于较好的表达不同词之间的相似关系。第四步模型训练，包括传统的有监督、半监督和无监督学习模型等，可根据应用需求不同进行选择。但在训练模型时可能会出现过拟合和欠拟合的状况。所谓过拟合就是学习到了噪声的数据特征，而欠拟合是不能较好的拟合数据。解决过拟合的方法主要有增加正则化项从而增大数据的训练量，解决欠拟合则要减少正则化项，增加其他特征项处理数据才行。第五步对建模后的效果进行评价，常用的评测指标有准确率(Precision)、召回率(Recall)、F值(F-Measure）等。准确率是衡量检索系统的查准率；召回率是衡量检索系统的查全率；而F值是综合准确率和召回率用于反映整体的指标，当F值较高时则说明试验方法有效。

3.2 自然语言处理基础研究

3.2.1 词法分析

词法分析主要包括分词、词性标注、命名实体识别和词义消歧。词性和词义标注是词法分析的主要任务。词性是词汇最基本的语法属性，使用词性标注便于判定每个词的语法范畴。词义标注、词义消歧主要解决多语境下的词义问题，因为在多语境下一个词可能会拥有很多含义，但在固定情境下意思往往是确定的。在中文自然语言处理的分词模块中，词法分析是最核心的一部分[4]，只有做好分词工作，剩下的工作才能顺利进行。命名实体识别的主要任务是识别文本中具有特定意义的词语如人名、地名等，并为其添加标注，是自然语言处理的一个重要工具。词法分析的实现主要通过基于规则、基于统计、基于机器学习的方法。

3.2.2 句法分析

句法分析的主要任务是为了确定句子中各组成成分之间的关系，也就是其句法结构，技术实现上主要分为修辞结构分析和依存关系分析，功能上可分为完全句法分析和局部句法分析。完全句法分析是要通过一套完整的分析过程获得一个句子的句法树，局部分析也叫浅层分析，仅获得局部成分的语法。目前应用较多的依存分析是指对句子中词汇之间的依存关系进行分析。

对完全句法分析来说，Chomsky形式文法是极为重要的理论，根据重写规则分为4级[5]，分别是0型文法（无约束文法）、1型文法（上下文有关文法）、2型文法（上下文无关文法）和3型文法（正则文法）。这4种文法统称为短语结构语法。浅层句法分析可分为两个子任务：其一是识别和分析语块，其二是分析语块之间的依附关系。依存句法也称从属关系语法。一个依存关系可分为核心词和依存词[6]。核心词是一个句子的根节点，一个句子只有一个，它负责支配句子中的其他词，核心词一般与依存词之间存在着一定的关系，如主谓关系、动宾关系和并列关系等。

3.2.3 语义分析

对于不同的语言单位，语义分析有着不同的意义。在词的层面上，语义分析指词义消歧；在句的层面上指语义角色标注；在篇章的层面上指共指消解[7]。语义分析是目前NLP研究的重点方向。

3.2.4 语用分析

语用分析主要是把文本中的描述和现实相对应，形成动态的表意结构。语用分析有四大要素：发话者、受话者、话语内容和语境。前两者指语言的发出者和接受者；话语内容指发话者用语言符号表达的具体内容；语境指言语行为发生时所处的环境，主要有上下文语境、现场语境、交际语境和背景知识语境。

4 自然语言处理的技术领域

自然语言处理作为一个多学科交叉的研究领域，涉及到许多的研究和应用技术，包括信息检索、文本分类和自动文摘等。信息检索（IR）有两方面的任务[8]，第一是存储海量信息，第二是根据用户需求快速查找相关信息；文本分类是根据一套分类规则对文本进行自动分类的过程；情感分析是一种通过判断文本情感极性去表征文档的技术；机器翻译是通过计算机将一种语言翻译到其他语言；社会计算是采用互联网、大数据和机器学习等技术来研究社会问题，并寻找出一种合适的方法去解决问题。接下来将针对信息抽取和自动文摘做详细介绍。

4.1 信息抽取（IE）

信息抽取是将嵌入在文本中的非结构化信息提取并转换为结构化数据的过程，从自然语言构成的语料中提取出命名实体之间的关系，是一种基于命名实体识别更深层次的研究[9]。信息抽取的主要过程有三步：首先对非结构化的数据进行自动化处理，其次是针对性的抽取文本信息，最后对抽取的信息进行结构化表示[10]。信息抽取最基本的工作是命名实体识别，而核心在于对实体关系的抽取。

4.1.1 信息抽取的主要方法

近年来，随着互联网的普及和大数据技术的实用化，非结构化（自由文本）文本信息的分析技术已成为NLP的必然。传统的信息抽取的方法主要有两种：基于规则和基于统计的方法。早期主要采用基于规则的方法，但由于其自身的局限

性造成了一定的困扰，比如人工制定规则的过程较复杂也消耗人力，工作效率较低。所以后来基于概率的方法慢慢占据主要地位，虽然基于概率的方法可以在一定程度上弥补基于规则的方法的缺陷，终究也不是完美的，所以需要找到可以使两种方法相辅相成的方法，使信息抽取的效果更佳。近年来，信息抽取工作越来越依赖机器学习的算法，所以机器学习在一些方面的突破为信息抽取提供了技术上的支持[11]。

Golshan[12]提出该领域的最新方法有基于机器学习的方法和基于深度学习的方法，近年来基于深度学习的方法一直是研究者关注的焦点，所有的这些方法为信息抽取技术（IE）[13]的出现奠定了基础。Niklaus 等人[14]概述了解决 Open IE 的几种方法，并将它们归为 3 类：基于规则的、基于学习的、基于 clause 系统的方法。Cui[15]等人提出了一种基于编译码框架的神经 Open IE 方法，将 Open IE 转换为一个序列到序列生成的问题，其中输入序列是句子，输出序列是一种带有特殊占位符的元组。研究表明，神经 Open IE 系统的性能显著优于多数基线，它的精度和召回率方面也明显优于其他方法。

4.1.2 信息抽取的主要工作

信息抽取主要工作包括实体识别与抽取、实体消歧、关系抽取和事件抽取等。其中基础性工作是命名实体识别（NER），其主要任务是识别文本中具有特定意义的词语，并为其添加相应的标注，为后续工作奠定基础。早期命名实体识别采用基于规则的方法，更多的是人工编写规则，这种方法准确率高召回率却很不理想，后来人们利用机器学习建立知识库再对文本进行处理的方法提高效率。国内近几年的研究热点集中在应用阶段，命名实体识别也进入到实用阶段。文本中每个实体可对应到多个真实世界实体中，实体消歧就是确定某一实体所指向的某一确定实体，主要有基于聚类的实体消歧和基于实体链接的实体消歧。关系抽取作为信息抽取的核心工作，主要任务是获取实体之间在语义上的联系。早期信息抽取主要用到的方法是模式匹配，后来又推出了基于词典驱动的方法，如今主要在基于本体的关系抽取的基础上，采用机器学习的方法来获取关系特征。

4.2 自动文摘

自动文摘是利用计算机按照某一规则自动地对文本信息进行提取、集合成简短摘要的一种信息压缩技术，旨在实现两个目标：首先使语言的简短，其次要保留重要信息[16]。

4.2.1 自动文摘的分类

从 1955 年 IBM 公司 Luhn 首次进行自动文摘的实验到现在的几十年，根据摘要方式[17]的不同主要分为两种：抽取式摘要和生成式摘要两种，抽取式摘要是选取原文中部分关键词组合成一篇摘要；生成式摘要是指当计算机通读原文并理解了文章的基础上，间接凝练出原文的主旨要点。除此之外还可以根据输入文本的数量分为单文本摘要和多文本摘要[10]等。

4.2.2 自动文摘生成方法

自动文摘的主要过程有三步，首先对语料进行预处理，识别冗余信息；其次是对文本内容进行选取和泛化；最后对文摘进行转换和生成，就是对文本内部进行重组生成文摘[16]，生成的摘要具有压缩性、内容完整性和可读性的特点。

自动文摘的主要方法包括：基于规则的方法、基于图模型的方法、基于理解的方法和基于结构的方法等。Lead 方法是基于规则的抽取式自动摘要中的常用方法[18]，虽然规则简单但是效果较好，特别是对于新闻类的文摘。图模型可直观表达出词与词之间的关联信息，弥补传统向量法的不足。经典的 TextRank 算法模型就属其中一种，用 G=(V,E) 来表示，其中 V 表示图中所有节点的集合，E 表示图中所有边的集合，是 V*V 的子集，图中任两点 V_i，V_j 之间的边权重为 w_{ji}，节点 V_i 的得分如公式 3-1 所示。

$$WS(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}WS(V_j) \quad (3\text{-}1)$$

其中 d 为阻尼系数，取值范围是 0 到 1，表示图中的某一节点跳转到其他任意节点的概率，一般取值为 0.85。除此之外，还可以使用 Word2vec 方法计算相似度，将各句子用向量表示，然后计算余弦距离并选择与原文语义相似度最高的句子作为文摘句[19]。

Neto[20]等人提出了一种基于可训练机器学习算法的摘要过程，实验表明采用朴素贝叶斯的可训练方法分类器明显优于所有基线方法。基于大量金融领域的长文本语料，王帅[21]提出了一种新摘要方法叫做：TP-AS，该方法采用两阶段自动生成摘要方法，其准确性在 ROUGE-1 的指标下达分别达到了 36.6%（词）和 33.9%（字符），结果明显优于其他方法。Liu[22]提出了一种基于模型的 NEXTSUM 方法，生成的摘要的长度与人工编写的黄金标准的长度呈正相关，表明可以隐式地捕获源文章中有多少值得摘要的内容。

5 自然语言处理的预测和展望

虽然自然语言处理的相关研究比较抽象，但其最基础的研究还是对语法、句法和语义的研究，关注的核心在于语言和文本。自然语言处理的难点在于理解语言不能光靠逻辑，还要有强大的知识库，需要有这些支撑才能更好的处理数据并对文本进行进一步的理解和分析。

从长远来看，自然语言处理具有广阔的应用领域和前景，作为一门由计算机科学、人工智能和语言学三科融合的新兴领域，它的长远发展对每个学科都具有重大的意义和影响力。未来自然语言的发展趋势可能从人工构建知识到自动构建，人们可以利用一些显性知识构建一种方法，挖掘语言成分之间的关系，这样就避免了人工的繁琐和耗时。在文本的理解和推理层面可以由浅及深，完成对文本的深层次理解。哈尔滨工业大学刘挺教授在第三届中国人工智能大会上提到：可以使阅读理解作为一个深入探索自然语言理解的平台，Google 也已经推出了这样的测试机，也就是说让计算机理解一篇文章，接下来人类对计算机进行提问，观察计算机的回答能力完成测试。

未来自然语言处理的发展趋势是 NLP 与许多领域的深度结合，从而为各相关行业创造价值。银行、电器和医学等领域对自然语言处理的需要都在日益提高，NLP+ 与各行业的结合越紧密，专业化的服务趋势就会越来越强。

参考文献：

[1] 冯志伟·自然语言处理的历史与现状[J]·中国外语,2008(01):14-2·

[2] 宋一凡·自然语言处理的发展历史与现状[J]·中国高新科技,2019(03):64-66·

[3] Zeroual I , Lakhouaja A · Data science in light of natural language processing: An overview[J]· Procedia Computer Science, 2018, 127:82-91·

[4] 邢蕾·英汉机器翻译中译文自动生成系统设计[J]·现代电子技术,2018,41(24):86-89·

[5] 郑伟发·汉语句法分析研究综述[J]·信息技术,2012,36(07):72-74+78·

[6] 张苗苗·汉语格库构建方法的研究[D]·北京交通大学,2018·

[7] 宋洋,王厚峰·共指消解研究方法综述[J]·中文信息学报,2015,29(01):1-12·

[8] 徐凡,朱巧明,周国栋·篇章分析技术综述[J]·中文信息学报,2013,27(03):20-32+55·

[9] 郭喜跃,何婷婷,胡小华,陈前军·基于句法语义特征的中文实体关系抽取[J]·中文信息学报,2014,28(06):183-189·

[10] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2013：486,719,726·

[11] 郭喜跃,何婷婷·信息抽取研究综述[J]·计算机科学,2015,42(02):14-17+38·

[12]Golshan P N , Dashti H A R , Azizi S , et al· A Study of Recent Contributions on Information Extraction[J]· 2018· J· Piskorski and R· Yangarber· Information extraction·

[13]J· Piskorski and R· Yangarber, "Information extraction: past, present and future," in Multi-source, multilingual information extraction and summarization, Springer, pp· 23-49, 2013·

[14]Niklaus C , Cetto M , Freitas, André , et al· A Survey on Open Information Extraction[J]· 2018·

[15]Cui L , Wei F , Zhou M · Neural Open Information Extraction[J]· 2018·

[16]Text summarization using Wikipedia[J] · Yogesh Sankarasubramaniam,Krishnan Ramanathan,Subhankar Ghosh· Information Processing and Management · 2014 (3)·

[17] 黄波,刘传才·基于加权 TextRank 的中文自动文本摘要[J/OL]·计算机应用研究:1-5[2019-04-18]·https://doi·org/10·19734/j·issn·1001-3695·2018·07·0528·

[18] 张洪荣·中文自动文摘关键技术研究与实现[D]·哈尔滨工业大学,2018·

[19] 李小涛,游树娟,陈维·一种基于词义向量模型的词语语义相似度算法[J/OL]·自动化学报:1-16[2019-04-19]·https://doi·org/10·16383/j·aas·c180312·

[20]Neto J L , Freitas A A , Kaestner C A A · Automatic Text Summarization Using a Machine Learning Approach[C]// Advances in Artificial Intelligence, 16th Brazilian Symposium on Artificial Intelligence, SBIA 2002, Porto de Galinhas/Recife, Brazil, November 11-14, 2002, Proceedings· Springer Berlin Heidelberg, 2002·

[21] 王帅,赵翔,李博,葛斌,汤大权·TP-AS: 一种面向长文本的两阶段自动摘要方法[J]·中文信息学报,2018,32(06):71-79·

[22]Liu J , Cheung J C K , Louis A · What comes next? Extractive summarization by next-sentence prediction[J]· 2019·

（收稿日期：2019-06-13）