《平凡的世界》小说文本分析

1、引言

在数字化、网络化、全球化的科技迅猛发展的时代下，互联网上人与人之间的沟通交流方式越来越来丰富多彩，产生的数据总量及其复杂度都呈现出“爆炸式增长”的模式。面对海量、结构复杂的文本数据，通过以往的人工浏览的方式从大量的文本中获取有效的信息变得十分困难。近些年来，自然语言处理(Natural Language Processing)逐渐兴起，融合了计算机科学，统计学，代数学，语言学等多门学科，其目的是为了让计算机学会认知、理解、运用自然语言，从而帮助人们快速、准确的获取蕴含在文本中的信息。这里的“自然语言”是指自然世界中存在的各种人类语言，如中文、英文、德语、日语等自然界中人们沟通交流所使用的语言，并非指c、c++、java、goland等计算编程语言。由于其高效，快速，准确的特性，自然语言处理广泛应用于信息检索(Information Search)、数据挖掘(Data Mining)、机器阅读理解(Machine Reading Comprehension)、机器翻译(Machine Translation)、情感分析(Sentiment Analysis)、文本分类(Text Classification)、语音识别(Speech Recognition)等领域。随着计算机的算力不断提升以及存贮容量的增加，自然语言处理取得的效果和准确率也在不断的提升，能够成为科研人员或者课堂老师分析文本数据的有效利器。

由于长篇小说《平凡的世界》的篇幅过长，其中的人物关系复杂，故事情节跌宕起伏，对其感兴趣的读者在阅读的过程中没有足够的时间时，难以将整本书读完。即便读过的读者也需要往复阅读这本小说，才能理清其中的故事脉络和人物关系网络。而《平凡的世界》这本小说其中所蕴含的人情事故的处事原则和人生哲理也将会使读者受益。为帮助《平凡的世界》的读者能够掌握其中的知识，且鉴于自然语言处理技术的强大，本文利用自然语言处理中较为流行的技术帮助对《平凡的世界》长篇小说感兴趣的读者进行文本分析，帮助其掌握小说的主题思想，故事情节发展线以及小说人物之间的关系网络。

2、国内外研究现状

文本分析属于自然语言处理下的一个重要领域，而自然语言处理属于深度学习和机器学习的范畴。本小节主要将自然言语的发展历程划分为5个阶段进行介绍。

2.1萌芽期（20世纪40年代-20世纪50年代末）

在第二次世界大战后，图灵(Turing)于1936年首次提出“图灵机”得概念。图灵模型的诞生带动了自动化的萌芽，这被认为是当今计算机科学发展的根基。图灵在自动化研究上的工作引领了同时期的科学家们在该领域的不断思考和研究。其中数学家沃尔特皮茨(Pitts)与医学博士麦卡洛克(McCulloch)的共同工作使得神经网络得最初模型问世；以及美国数学家、逻辑学家Kleene在有限自动机和正则表达式上的研究为后来的机器学习奠定了基础。1956年，香农(Shannon)将离散马尔可夫过程的概率模型应用在描述语言的自动机上。1956年，语言学家乔姆斯基（Chomsky）博士在香农的工作上，提出使用有限状态机来识别句法以及定义有限状态机过程产生的语言为有限状态语言，包括上下文无关语法。这些工作形成了前期机器语言处理的理论基础。

该时期，基于概率算法的语音和语言处理也得到了发展，香农提出了在信道中传输的信息的编解码方式以及噪声的存在对信道的干扰。香农将热力学中的“熵”的概念应用到统计信道中信息容量的大小，开创了将概率方法应用到英文信息测度领域的先河。

1952年，贝尔实验室创建了第一个基于统计的语音辨识系统，该系统通过说话人的语音来辨识输出0到9之间的数字。在特定的条件下，该系统能够达到97%至99%的正确率（Davis et al., 1952）。

2.2 快速发展期（1957年-1970年）

20世纪50年代末至60年代初，自然语言处理形成了两种研究范式，符号派研究学者认为应当将基于规则的符号表用于NLP中的语言处理，而随机派学者则主张使用概率方法进行研究。以乔姆斯基为代表的学派在形式语言理论和生成句法上进行不断研究，形成了一套解析算法理论，由最初的由上至下以及自下而上到动态编程技术。1958年至1959年，哈里斯（Harris，1962）在宾夕法尼亚大学的实验室创造了最早的完整解析系统Transformations and Discourse Analysis Project（TDAP）；以香农为代表的学术流派提出了人工智能的概念，他们注重研究推理和逻辑问题的研究。在这一时期，科研人员将模式匹配（Patten Matching）与关键词查找等启发式的研究方法应用在问题回答（Question Answeing）领域上，建立了自然语言理解系统的最初模型。

20世纪50年代末，贝叶斯方法理论开始被应用到光学字符识别问题上。布莱索（Bledsoe）和布朗宁（Browning）建造了一个基于贝叶斯理论的文本识别系统，该系统在一个项目庞大的词典中查找给定的词概率，并将它们乘起来计算特定字符序列出现的概率。同时期形成一些可以用于科学研究的语料库。

2.3 低速发展期（1971年-1993年）

由于语料库的缺乏，以及计算机的算力不高，当时自然语言处理的应用不能取得较好的结果。

但相关的研究仍然在继续着，进行随机范式研究的学者提出了隐式马尔科夫模型（Hidden Markov Model，HMM）；基于逻辑推理的研究学者提出了变形语法(Metamorphosis grammar )、固定子句语法（Definite clause grammar）、功能语法（functional grammar）和词汇功能语法（lexical functional grammar）等语法规则；自然语言理解中诞生了SHRDLU系统，它能够接受自然语言的文本命令并作出相应的行动。基于逻辑推理的范式与自然语言理解范式相结合，推动了语义表示的发展，譬如LUNAR问答系统的诞生。

2.4 复苏融合期（1994年至今）

这一时期的科学家们的研究方式逐渐从以经验为导向向数据驱动学习过渡，着得以与计算机的计算能力的大幅度提高。各种研究方式相互借鉴，汲取他人所长，概率方法论与数据建模模型的结合使得信息提取有了较好的效果。

这一时期，NLP研究人员可以在互联网上获取种类丰富，数据量庞大的预料库来训练自己的模型。统计机器学习社区的活跃度保持较高的水平：支持向量机的提出（Boser et al.，1992；Vapnik，1995）、最大熵技术以及多项式回归（Berger et al.，1996）、贝叶斯图论模型（Pearl，1998）、非监督统计方法受到广泛的关注。

下面是21世纪以来应用在NLP领域具有里程碑的重点技术

A、神经语言模型（Neural language models）。该模型旨在根据给定的文本中的词语或前n个词语预测下个词语的可能情况，将概率最大的词语作为下一个词语输出（Bengio et al.，2001）

B、多任务学习（Multi-task learning）。该模型是为了解决单一预料库不足导致训练样本不充分的问题，它将多个任务一起训练，共享任务模型之间的参数，能够较好使每个任务模型得到较好的训练（Collobert，2008）

C、词嵌入（Word embeddings）。Mikolov认为每个词语不是单独存在的，每个单词的意思依赖其具体的上下文，同时词语的特征表示层应当具有普遍性，不应当随具体的研究领域有剧烈的变化。Mikolov提出该方法使得每个单词都映射到对应的特征层中的一个点，并将其作为下游任务的输入（Mikolov，2013）

D、用于自然语言处理的神经网络（Neural networks for NLP）。NLP中流行的神经网络包括卷积神经网络（Convolutional neural networks）其通过输入-编码-激活-卷积-池化-归一化等操作输出期望值、循环神经网络（Recurrent neural networks）该网络能够记住一定长度的句子信息，避免局部局限的问题。

E、序列到序列模型（Sequence-to-sequence models）。该模型用于机器翻译，对话代理等领域，取得较好的实际效果（Sutskever，2014）

F、注意力机制（Attention）。该机制能够学习较长范围的语句信息，得到比较综合全面的文本信息，适合于阅读理解等方向（Bahdanau，2015）

G、预训练语言模型（Pretrain language models）。由于现在的网络模型的庞大与学习的复杂性，且认为一般问题遵循基本的表示规律，科研人员提出了预训练方法，用于学习输入的特征表示，将其作为下游任务的输入，并且根据特殊问题进行微调（fine-tuning）能够缩短训练时间和降低训练复杂度，且取得较好的实验结果（Conneau et al.，2017；McCann et al.，2017；Subramanian et al.，2018）

理论模型的蓬勃发展，促进了相关软件的诞生，帮助人们将注意力放在行业务领域问题的分析上，减轻了人们的学习和应用负担。

1. 研究展望

本文希望通过运用文本挖掘中的技术如整体词频、词云的展示获取《平凡的世界》小说文本中的前50个关键词信息，对各个章节进行文本摘要，根据得出的总结能够准确得到每一章的中心思想，以及小说故事情节的发展脉络和小说人物关系网络分析，形成知识图谱。

1. 参考文献

[1]Daniel Jurafsky, James H.Martin.2012.Speech and Language Processing[M]:Stanford

University, University of Colorado at Boulder

[2]陈瑛.2021.基于Python的《三国演义》人物关系网络的构建与分析[D].广州:广州工程技

业学院信息工程学院

[3]宋冠仪.2021.基于bert的多任务文本分析研究[D].山东:山东大学

[4] Roman Collobert, Jason Weston, Michael Karlen, et al.2011.Natural Language Processing

(almost) from scratch.arXiv:1103.039v1

[5]Tomas Mikolov, Kai Chen, Greg Corrado.2013.Efficient Estimation of Word Representations in Vector Space.arXiv.1301.3781v3

[6]Yoshua Bengio, Rejean Ducharme, Pascal Vincent, et al.2003.A Neural Probabilistic Language Model.Journal of Machine Learning Research 3 1137-1155

[7]Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al.2019.Bert:Pretraining of Deep Bidirectional Transformers for language Understanding.arXiv:1810.04805v2

[8]余凯，贾磊，陈雨强等.2013.深度学习的昨天、今天和明天[J.].计算机研究与发展:ISSN 1000-1239

[9]Minjoon Seo, Aniruddha Kembhavi, All Farhadi, et al.2018.Bi-Di-directional Attention Flow For Machine Comprehension.arXiv:1611.01603v6

[10]Rada Mihalcea and Paul Tarau.2017.TextRank:Bringing Order into Text.Department of Computer Science University of North Texas

[11]Changhan Wang, Anirudh Jain, Danlu Chen, et al.2019.ViSeq:A Visual Analysis Toolkit for Text Generation Tasks.arXiv:1909.05424

[12]Sunny Lai, Kwong Sak Leung and Yee Leung.2018.SUNNYNLP at SemEval-2018 Task 10:A Support-Vector-Machine-Based Method for Detecting Semantic Difference using Taxonomy and Word Embedding Features.Department of Computer Science and Engineering, Department of Geography and Reasource Management

[13]Christos Louizos, Kevin Swersky, Yujia Li, et al.2017.The Variational fair Autoencoder.arXiv:1211.00830v6

[14]Yaroslav Ganin, Hana Ajakan, Hugo Larvochelle, et al.2016.Domain-Adversarial Training of Neural Networks.arXiv:1505.07818v4

[15]Sebastian Ruder and Barbara Plank.2018.Stong Baselines for Neural Semi-supervised Learning under Domain Shift.arXiv:1804.09530v1