人工智能在自然语言处理中的演进与应用

蒋添爱

2022 年 12 月 30 日

摘要：自然语言处理是人工智能领域中十分重要的一个研究方向，其发展历程经过了新生期、复兴期和飞跃期三个阶段。自然语言处理在日常生活中的各方各面也有着极其重要的应用，如利用 word2vec 模型、支持向量机分类模型进行舆情分析。

关键词：人工智能、自然语言处理、舆情分析

自然语言处理是人工智能领域中十分重要的一个研究方向，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，其作为人类与机器之间沟通的桥梁，以实现人机交流的目的。[8]

1 自然语言处理的发展历程

随着人工智能技术的不断发展，自然语言处理也随之更新迭代，其发展历程与人工智能是保持高度相似的，也先后经过了新生期、复兴期和飞跃期，人工智能的三个阶段：规则推理阶段、机器学习阶段和深度学习阶段也均在自然语言处理领域上有所探究发展，提出了相应的方案理论。

1.1 新生期

在规则学习阶段中，由于机器翻译的社会需要，以及图灵测试中机器对提问者问题的理解，自然而然地诞生了自然语言处理技术。此时是自然语言处理的基础研究阶段，理论方法也是基于规则的。然而类似于当时所有基于规则学习方法的人工智能领域，其所面临的不可避免的缺点首先即是规则不可能覆盖所有语句，其次这种方法对开发者的要求极高，开发者不仅要精通计算机还要精通语言学，因此，这一阶段虽然解决了一些简单的问题，但是无法从根本上将自然语言理解实用化。在这种被称为”飞鸟派”方法的约束限制下，自然语言处理技术也在一段时期内陷入了沉寂。

1.2 复兴期

而当人工智能技术步入机器学习阶段后，自然语言处理也转而采用基于统计的方法。随着互联网的高速发展，语料库的日益丰富以及硬件不断更新完善，自然语言处理思潮由理性主义向经验主义过渡，基于统计的方法逐渐代替了基于规则的方法，最终使准确率有了质的提升。机器学习算法是一类从数据中自动分析获得规律，并且利用规律对位置数据进行预测的算法，其与推断统计联系尤其密切，仅仅是利用大量的数据，就可以让计算机自己学习，大大降低了人工智能相关问题的开发难度 [6]。基于统计、基于实例和基于规则的语料库技术在这一时期开始蓬勃发展，各种处理技术开始融合自然语言处理的研究再次繁荣。

1.3 飞跃期

深度学习的概念和方法出现之后，自然语言处理又有了突飞猛进的变化。深度学习是一种将原始数据通过一些简单但是非线性的模型转变成更高层次、更加抽象表达的特征学习方法，一定程度上解决了人类处理“抽象概念”这个亘古难题。深度学习方法具有强大的特征提取和学习能力，可以更好地处理高维度稀疏数据，因此在 NLP 领域中取得了长足发展 [7]。通过卷积神经网络 (Convolutional Neural Network, CNN) 和递归神经网络 (Recurrent Neural Network, RNN)，以及后续的注意力机制、预训练模型等，实现了如语音识别、问答系统、情感分类、上下文预测等自然语言处理上的商业化应用。

尽管人工神经网络在数学上是完美的，可是它的运行机理却不容易从语言学和普通常识的角度得到理性的解释，而且人工神经网络也难以应用人类在长期的语言学研究中积累起来的丰富多彩的语言知识来改进自身的不足。在自然语言处理研究中，在大力推广人工神经网络的经验主义方法的同时，也要逐步复兴近年来受到冷落的理性主义方法，让神经网络从语言的理性规则中吸取营养，不断完善，增强它的可解释性。基于语言数据的经验主义方法一定要与基于语言规则的理性主义方法结合起来，才是自然语言处理的必由之途，也是当代人工智能研究的金光大道 [4]。

2 自然语言处理的应用

对于日益成熟的人工智能以及自然语言处理领域来说，其在日常生活中的各方各面也有着极其重要的应用。在机器翻译方面，随着机器学习和深度学习的兴起，机器翻译领域迎来了新的机遇，翻译质量也有了质的飞跃，Google、百度等公司都已经将线上机器翻译系统升级到神经网络翻译模型，每天为数亿用户提供高质量服务；在问答、对话系统中，人工智能和自然语言处理技术也使得该类系统能逐步变得更加”智能”，具有实用价值，如苹果公司的 Siri、微软的小娜，最近兴起的 ChatGPT 均可以看见这类技术的应用；在舆情分析中，自然语言处理则能将现实中各种舆情数据转化为计算机能理解分析的数据形式，并按照用户需求输出用户所关心的信息，从而把控社会舆论的发展走向。

以舆情分析为具体例子，在数据收集过程中，可以采用自然语言处理中的分词程序和 word2vec 模型对数据进行处理。分词程序用于对收集到的语料库即文本数据进行分词和词性标注，其中词性标注则为数据打标签，如布朗语料库 (Brown Corpus) 中就将各种单词归类为名词 (NN)、动词 (VB)、副词 (RB) 等 [2]。word2vec 模型则是将自然语言中的单词转化为词向量的一种方法。词向量是一种由神经网络学习得到的词的分布式表示，通过 Word2Vec 训练出的各词的词向量提高了词相似性的计算结果 [3]，即 word2vec 模型的输入是收集的文本语料库，输出是一组向量：该语料库中所有单词的特征向量。根据该特征向量便能反映单词的语义信息，语义上相似的单词在空间内距离即向量很近，故可以在词向量中发现类似于”国王-男性 = 皇后-女性”这样有趣的现象 [1]。

通过 word2vec 将文本转换为深度神经网络可以理解的数字形式后，便可以从候选特征向量中基于情感词典筛选出有效特征，再利用如支持向量机 (Support Vector Machines, SVM) 分类模型对词向量进行二分类，将词语划分为积极的和消极的两类，从而得到词语的舆情指标 [5]。该模型属于机器学习的一种模型，其学习策略可用”最大化数据集与分离超平面的几何间隔”或”最小化合页损失函数”，从而求得最优化的参数。对于模型效果的评估则可用混淆矩阵，根据混淆矩阵中的精确率、召回率等指标得到该模型的分类效果。

参考文献

[1] Xin Rong. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738, 2014.

[2] Wikipedia contributors. Brown corpus — Wikipedia, the free encyclopedia, 2022. [Online; accessed 30-December-2022].

[3] Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[4] 冯志伟. 神经网络, 深度学习与自然语言处理. 上海师范大学学报：哲学社会科学版, 50(2):13, 2021.

[5] 张冬雯, 杨鹏飞, and 许云峰. 基于 word2vec 和 svmperf 的中文评论情感分类研究. 计算机科学, (S1):5, 2016.

[6] 李彦峰. 人工智能在自然语言处理中的应用. 襄阳职业技术学院学报, 17(4):5, 2018.

[7] 罗枭. 基于深度学习的自然语言处理研究综述. 智能计算机与应用, (4):5, 2020.

[8] 赵京胜, 宋梦雪, and 高祥. 自然语言处理发展及应用综述. 信息技术与信息化, (7):4, 2019.