文章编号：2096-1472(2021)-09-09-04
DOI:10.19644/j.cnki.issn2096-1472.2021.09.002

人工智能自动生成海上大风预报报文研究

简 俊[1]，王 衡[1]，孙 正[1]，吴冠霖[1]，苏 欣[1]，陈三君[2]

(1.大连海事大学，辽宁 大连 116026；
2.中海石油(中国)有限公司深圳分公司，广东 深圳 518067)
✉jianjun@dlmu.edu.cn; wheng@dlmu.edu.cn; szheng@dlmu.edu.cn; wuguanlin@dlmu.edu.cn;
sux@dlmu.edu.cn; chenshj3@cnooc.com.cn

摘 要：通过人工智能领域中的自然语言工具将不直观的专业海上风速预报数据转化为友好易懂的分析文本，提高气象预报服务的时效性和合理性。首先提取中国海洋石油公司南海东部四个海上区块的原始预报数据，利用神经网络对数据进行回归拟合，得到一条显示未来风速上升下降的单调区间、极值点和增减幅度的光滑曲线，再利用自然语言处理基于规则和基于统计相结合的方法自动生成预报文本。该方法可将现行需要人工45—60 分钟才能完成的预报报文减少到2—4 分钟内完成，大大提高了海洋气象预报的工作效率。

关键词：海上大风预报；自动生成；自然语言处理；神经网络拟合
中图分类号：TP39 **文献标识码**：A

Study on Automatic Generation of Marine Wind Forecast Text by Artificial Intelligence

JIAN Jun[1], WANG Heng[1] , SUN Zheng[1], WU Guanlin[1] , SU Xin[1] , CHEN Sanjun[2]

(1. *Dalian Maritime University*, *Dalian* 116026, *China;*
2. *Shenzhen Branch, CNOOC* (*China*) *Co*., *Ltd*., *Shenzhen* 518067, *China*)
✉jianjun@dlmu.edu.cn; wheng@dlmu.edu.cn; szheng@dlmu.edu.cn; wuguanlin@dlmu.edu.cn;
sux@dlmu.edu.cn; chenshj3@cnooc.com.cn

Abstract: This paper proposes to apply Artificial Intelligence-Natural Language tool to transform professional and non-intuitive meteorological data into understandable and user-friendly texts, thus improving the speed and rationality of meteorological services. The raw meteorological forecast data of the CNOOC's four offshore blocks in Eastern South China Sea is retrieved firstly, and the neural network is applied to perform regression fitting on the data, so to generate a smooth curve showing the monotonic interval, extreme points, and increase/decrease range of future wind speed. Then, Natural Language tool is applied to process the forecast texts which are automatically generated based on a combination of rules and statistics. Using the proposed method, the current 45-60 minutes manual forecasting work could be completed within 2-4 minutes, which greatly improves the efficiency of sea weather forecasting.

Keywords: sea wind forecast; automatic generation; natural language processing; neural network fitting

1 引言(Introduction)

海洋气象预报是针对海上或近岸作业单位所处的海洋环境的气象情报，用以保证海洋作业的安全及在可能的条件下提高海上作业的效率。提高航运安全保障技术中的一个重要内容是精细化的海上风力预报与灾害性大风浪预警，精细化既包括面向对象，也包括准确性和时效性[1]。近年来，海洋经济的高速发展对海洋气象工作的全面融入提出了新的更高的要求，因此基于传统的天气学和统计学的预报方法较难适应

基金项目：国家自然科学基金国际(地区)合作交流项目“中南半岛和华南极端降水事件年代际变化的机理、影响和未来预估”(41861144014)；国家自然科学基金杰出青年科学基金项目“海气自然变率影响青藏高原热源的过程和机理”(41725018)。

www.jgczz.com

海洋气象预报发展的现状，当前的海洋气象预报准确率和众多的预报要素及预报的内容需求越来越依赖于数值模式技术的发展。同样是基于气象数据，传统人工分析预报要耗费大量的预报人力和物力，准确性和精确性亦有待提高。而在大数据时代，使用全球大气环流模式产品和人工智能中的自然语言生成算法进行气象水文自动预测和台风预报已成为当今海洋气象预报发展的需求[2]。

2 数据处理(Data processing)

2.1 气象水文预报数据及人工分析报文

从欧洲中期天气预报中心(European Centre for Medium—Range Weather Forecast，ECMWF)的数值预报产品提取中国南海东部四个海上油气作业区块陆丰、流花、西江、恩平(下文分别用LF、LH、XJ、EP代替)未来10 天的原始预报数据，包括起报时间、目标时间、预报提前量、风速、风向、平均浪高和单次最大浪高等。

在专业教师的指导下，航海气象小组成员每天基于上面的预报数据和近期西风带槽脊变化等大尺度大气环流形势进行分析，写出自己的气象预报报文，并利用Excel软件做出风浪变化趋势图，再把结果发给南海东部四个区块的海上平台作业人员及陆上协调管理人员。此工作自2019 年初开始共持续600余天，每天坚持工作45—60 分钟生成一份预报报文，有效地协助我国相关企业安排海上作业时间窗口，规避恶劣天气造成的经济损失，但在产生较大经济效益的同时，也耗费了大量人工。

2.2 邮件收发

由于船舶及海洋平台上的通讯条件有限，通过Inmarsat海事卫星通信系统[3]虽然可以接入互联网，但费用高、带宽小，管理级人员大多数时候以收发电子邮件获取信息为主，不能像在陆地上那样比较方便地使用其他网络终端[4]。因此从方便用户及节省费用角度出发，设计采用云端服务及电子邮件交互的方法访问、接收、处理、发送各类数据信息，即大多数工作在校园内的服务器上自动进行，包括把原始预报数据下载到本地，通过编程语言处理后，再将最终报文结果发给海上用户等，全程无人工干预。

在服务器端自动收取邮件最常用的程序工具是POP3[5]协议和Python语言中的Zmail模块，经过试验对两种方法的优劣对比如表1所示。

表1 POP3协议和Zmail模块对比

Tab.1 Comparison of POP3 protocol and Zmail module

POP3协议	Zmail模块
存在服务端拒信的问题	自动填充大多数导致服务端拒信的头信息
需要找寻服务商的smtp地址、端口号	自动寻找服务商端口号地址，自动选择合适的协议
解析邮件复杂，各种格式乱码错误	只依赖于Python 3，有利于嵌入其他项目

Python标准库提供了smtp模块，用于实现SMTP协议，发送邮件，但SMTP协议存在与POP3协议同样的问题。而Zmail模块具有邮件发送功能，且代码简单不易出错，故邮件收发皆使用Zmail模块实现。

2.3 熵值法综合处理多源气象信息

以2020 年9 月7 日的工作为例，首先对通过Zmail模块获取的南海东部LF、LH、XJ、EP四个区块2019 年1 月4 日—2020 年9 月6 日的600多封邮件进行分析，转化为DataFrame格式，并对当天的原始起报数据进行清洗，删除异常值和空缺值，得到表2中的数值。表2中数据从左到右分别代表(北京时间)年、月、日、时、风速(m/s)、风向(度)、平均浪高(m)、单次最大浪高(m)，从上到下第二行开始为每隔6 小时步长的预报结果。

表2 2020 年9 月7 日原始数据清洗后的结果

Tab.2 Original data after cleaning on September 7, 2020

属性 / 提前量	0	1	2	3	4	5	6	7
1	2020	9	7	20	6.28	143.4	0.9	1.4
2	2020	9	8	2	8.56	150	1.1	1.8
3	2020	9	8	8	6.88	152.5	1.3	1.9
4	2020	9	8	14	4.96	187.5	1.2	1.8
5	2020	9	8	20	2.56	128.2	1.2	1.7
6	2020	9	9	2	5.56	152.2	1.2	1.7

在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小；信息量越小，不确定性越大，熵也越大。根据熵的特性，可以通过计算熵值来判断一个事件的随机性及无序程度，也可以用熵值来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的影响越大[6]。假设数据中有 n 个样本 m 个指标，其中 X_{ij} 表示第 i 个样本第 j 个指标($1\leqslant i\leqslant n, 1\leqslant j\leqslant m$)。熵权法赋权步骤如下：

步骤1：通过公式(1)计算出数据标准化值。通常应用最大最小标准化方法对数据进行标准化的操作，将各指标由绝对值变为相对值且消除量纲对结果的影响。

$$Y_{ij}=\frac{X_{ij}-\min(X_i)}{\max(X_i)-\min(X_i)} \tag{1}$$

步骤2：通过公式(2)和公式(3)计算出各指标的信息熵。

$$p_{ij}=Y_{ij}/\sum_{i=1}^{n}Y_{ij} \tag{2}$$

$$E_j=-\ln(n)^{-1}\sum_{i=1}^{n}p_{ij}\ln p_{ij} \tag{3}$$

如果 $p_{ij}=0$，则定义 $\lim\limits_{p_{ij}\to 0}p_{ij}\ln p_{ij}=0$。

步骤3：通过公式(4)确定各指标权重，根据信息熵的计算公式计算出各个指标的信息熵为 $E_1,E_2,\cdots,E_m$，通过信息熵计算

www.rigzz.com

各指标的权重。

$$W_j=\frac{1-E_j}{m-\sum E_j}(0\leqslant j\leqslant m) \tag{4}$$

因此，当判断出海上不同点位的风速平均值、方差、中位数时，熵值法可以依据这几个特征把气象数据综合起来，计算得到四个区块(LF、LH、XJ、EP)对应的权重，如表3所示。

表3 各区块风速平均值、方差、中位数和权重表

Tab.3 Wind speed average, variance, median and weight table of each block

区块名称	LF	LH	XJ	EP
平均值	10.96	11.05	10.19	10.43
方差	3.74	3.79	3.40	3.43
中位数	11.32	11.32	10.51	10.60
权重	0.25177	0.25207	0.24774	0.24842

2.4 神经网络拟合

尝试使用随机森林[7]、神经网络[8]及支持向量机回归拟合所得风速数据。由于支持向量机回归数据过多会泛拟合，随机森林特征较少，都不适用于该数据，而神经网络可以用作小样本和大样本，且结果与特征数目无关，最终选择sklearn中神经网络的MLPRegressor[9](多层感知器)进行回归拟合，激励函数使用“Relu”(整流后的线性单位函数)，返回$f(x)=\max(0，x)$。然后选取最新一封邮件，利用熵值法算出四个海上作业区块9月7日到9月17日的平均数据，便于之后的MLPRegressor进行处理，得到如图1所示的南海东部未来10天风速预报图(虚线曲线表示熵值法，实线曲线表示神经网络拟合)。

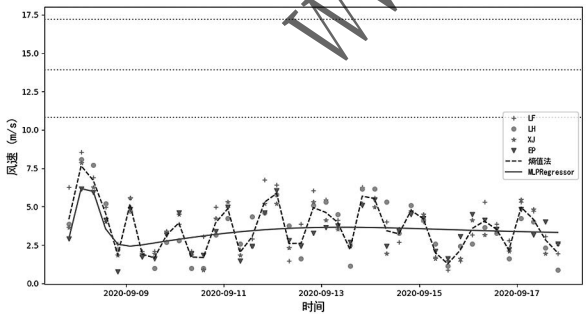

图1 南海东部未来10天风速预报

Fig.1 Wind speed forecast for the next 10 days in the Eastern South China Sea

3 基于规则的自然语言文本生成(Rule-based natural language text generation)

3.1 基本原理

自然语言有意义的基本单元是词，按照一定的句法规则将词组织在一起就成为句子，再由句子组成段落，由段落构成篇章。自然语言处理的基础研究主要包括词法分析、句法分析、语义分析、语用语境与篇章分析等。

自然语言生成(Natural Language Generation，NLG)是自然语言处理领域一个重要的组成部分，实现高质量的自然语言生成也是人工智能迈向认知智能的重要标志。作为人工智能和计算语言学的子领域，自然语言生成从抽象的概念层次开始来生成文本[10]。按照输入信息的类型划分，自然语言生成可以分为三类：文本到文本生成、数据到文本生成和图像到文本生成[11]。

3.2 中文分词技术

随着NLG技术的日益成熟，开源实现的分析工具越来越多，如Ansj、盘古分词等。分词工具选择了基于规则和基于统计的Jieba分词用于分词和关键词提取[12]。Jieba分词提供了精确模式、全模式和搜索引擎模式三种分词模式，例如Sample=‘根据最新气象数据显示，南海东部各平台风力等级波动不大！’运行结果如下：

全模式：根据/最新/新气象/气象/数据/显示/，/南海/海东/东部/各/平台/台风/风力/等级/波动/不大/！

精确模式：根据/最新/气象/数据/显示/，/南海/东部/各/平台/风力/等级/波动/不/大/！

搜索引擎模式：根据/最新/气象/数据/显示/，/南海/东部/各/平台/风力/等级/波动/不/大/！

3.3 关键词提取

关键词是代表文章重要内容的一组词，将已经向南海海上作业平台发送的600多份由气象小组成员编辑的气象分析报文作为语料库，提取出10个关键词。关键词提取算法步骤为：加载数据集→加载停用词表→数据集分词→过滤干扰词→训练算法[13]。关键词提取算法常用的有TF-IDF、TextRank、LSI和LDA四种，分别对四种算法进行训练，结果如下：

TF-IDF模型结果：

气象/风力/平台/冷空气/逐渐/迅速/西南/台风/趋势/风速/

TextRank模型结果：

风力/气象/平台/台风/风速/增大/逐渐/中心/西南/增强/

LSI模型结果：

阵风/台风/风力/平台/海区/天气/风速/迅速/南海/阵雨/

LDA模型结果：

随后/最大风速/季风/海浪/条件/逐渐/气压/百帕/速度/阵风/

www.rjczz.com

3.4 整体规则

根据人工气象预报文本格式，结合语料库提取的关键词制定预报文本模板。以熵值法曲线表示未来10 天风速总体增降趋势，以神经网络拟合曲线表示单调区间、极值点。标注7 级及以上大风为特殊点，需要特别关注，6 级及6 级以下风力仅一般性关注。气象预报文本生成流程如图2所示。

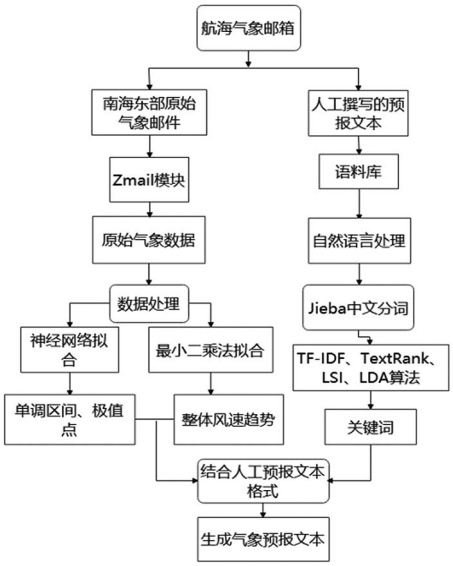

图2 自然语言生成规则流程图

Fig.2 Flow chart of natural language generation rules

3.5 文本生成

根据以上规则生成的南海东部LF、LH、XJ、EP各区块在2020 年9 月7 日生成的未来10 天内(最远覆盖到17 日)的预报文本如下：

“根据最新气象数据显示，未来10 天南海东部各区块风速有减弱的趋势，最大风力5 级，在8 日达到最大风速8.56 m/s。各区块风力等级都在6 级以下且波动较小，无7 级大风，有利于海上平台作业。

请关注最新气象预报！”

最后，利用Zmail模块将自动生成的气象预报文本通过电子邮件自动发送到南海东部的LF、LH、XJ、EP各平台应急管理负责人。经检验，全部工作可以在2—4 分钟内自动完成。

4 结论(Conclusion)

人工智能-自然语言生成工具已经广泛地应用到各行各业，产生了很大效益，气象报文的生成也不例外。从手工处理气象数据到预报文本的自动生成，具有很大的研究意义，同时实用性很强。目前该方法还有一些明显不足，如对四个区块只给出统一的报文文本及变化曲线，语言略生硬，偶尔出现遗漏或过度警报的问题，仍需要人工检查和纠正等。未来应当更新完善气象语料库，制定愈加人性化的文本生成规则，为提高海上气象预报技巧和效率做出贡献。

参考文献(References)

[1] 丁建军,罗兵,赵光平,等.精细化预报订正平台设计[J].气象,2008(11):89-95.

[2] 简俊,WEBSTER P J,朱大海.一种定量化的海面定点风速自动预报系统[J].航海技术,2012(05):9-11.

[3] 骆乐,包少彬,高恒伟,等.宽带海事卫星通信系统的技术分析[J].集成电路应用,2021,38(02):128-129.

[4] 徐丹.基于卫星网络的海事邮件通信系统架构设计[J].数字通信世界,2019(07):4-7.

[5] BELING H. Python使用POP3获取邮件信息[EB/OL].(2020-01-01)[2021-06-19]. https://blog.csdn.net/Beking17113/article/details/103797133.

[6] 王卓,高丛.基于信息论的熵值法的算法改进——以陕西省环境规制强度评价为例[J].西安石油大学学报(社会科学版),2016,25(01):22-26.

[7] PATIL D, RAJ R, SHINGADE P, et al. Feature selection and classification employing hybrid ant colony optimization/random forest methodology[J]. Combinatorial Chemistry & High Throughput Screening, 2009, 12(5):507-513.

[8] 吕砚山,赵正琦.BP神经网络的优化及应用研究[J].北京化工大学学报(自然科学版),2001(01):67-69.

[9] GIULIANO A. Using phidelta diagrams to discover relevant patterns in multilayer perceptrons[J]. Scientific Reports, 2020, 10(1):21334-21334.

[10] REITER E. Building natural-language generation systems[J]. Computational Lingus, 1996, 27(2):298-300.

[11] 李雪晴,王石,王朱君,等.自然语言生成综述[J].计算机应用,2021,41(05):1227-1235.

[12] 余洋.基于关键词的自动短文生成[D].武汉:华中师范大学,2020.

[13] 涂铭,刘祥,刘树春,等.Python自然语言处理实战:核心技术与算法[M].北京:机械工业出版社,2018:53-55.

作者简介：

简　俊(1976-)，男，博士，副教授.研究领域：海洋气象灾害预报，安全与保障.

王　衡(1993-)，男，硕士生.研究领域：海上风浪趋势预报.

孙　正(1998-)，男，硕士生.研究领域：风浪信息自动读取.

吴冠霖(1999-)，男，本科生.研究领域：自然语言生成.

苏　欣(2000-)，女，本科生.研究领域：交通信息安全.

陈三君(1983-)，男，本科，高级工程师.研究领域：海洋石油海上生产与应急管理.

www.rjgcz.com