气象落区文本自动生成研究

吴焕萍[1],吕终亮[2],张华平[3],罗 兵[2],高 健[3],李笑侃[3],何国豪[2],王永超[4]
WU Huanping[1], LV Zhongliang[2], ZHANG Huaping[3], LUO Bing[2], GAO Jian[3], LI Xiaokang[3], HE Guohao[2], WANG Yongchao[4]

1.国家气候中心,北京 100081
2.国家气象中心,北京 100081
3.北京理工大学,北京 100081
4.中国地质大学,北京 100083
1.National Climate Center, Beijing 100081, China
2.National Meteorological Center, Beijing 10081, China
3.Beijing Institute of Technology, Beijing 100081, China
4.China Universities of Geosciences, Beijing 100083, China

WU Huanping, LV Zhongliang, ZHANG Huaping, et al. Text generation on weather falling area description. Computer Engineering and Applications, 2014, 50(13):247-251.

Abstract: The text content needs the high quality characteristic of precision, efficiency, logicalness and as well as natural language expression in order to generate text description automatically for weather forecast and service fields, however currently it lacks more mature technique ways to solve all problems. Based on the analysis of the content of "weather public report" documents, it proposes a framework and its' processing of text generation on weather falling area description by introducing of Geographic Information Science(GIS) and Natural Language Processing(NLP) science. Feather more, it focuses on the four key issues for further discussion, such as information extraction and conceptual model building from massive history documents, geospatial area partitioning, meteorological elements spatial-temporal reasoning, text organization and generation and post-processing. Meanwhile, it proposes the technique implementation in detail for these issues. Compared with the content between forecaster manuscript and computer text, the results show the general text generation method has a potential application prospect in given meteorology forecast and services field.

Key words: Natural Language Processing(NLP); text feature extraction; meteorological data spatial analysis; text auto-generation

摘 要: 面向天气预报和气象服务的文本内容的计算机自动或者半自动生成方法,对文本生成质量要求较高,即要准确、高效、合理,还需要符合自然语言表达,存在较多技术问题。在深入分析中央气象台每日发布的"天气公报"文本内容的基础上,结合地理信息科学和自然语言处理科学方法提出了面向气象落区文本语言生成的基本原理与流程,重点从历史文本内容分析与特征提取、地理区域划分、气象要素空间分析、文本组织与生成等关键技术问题进行了深入讨论,并给出了相应的技术实现。计算机自动生成结果与预报员人工撰写的文本内容对比分析也较好地证明了面向特地领域的文本生成方法具有较好的应用前景。

关键词: 自然语言处理; 文本特征提取; 气象数据空间分析; 文本自动生成
文献标志码: A **中图分类号:** TP39 doi:10.3778/j.issn.1002-8331.1208-0464

基金项目: 中国气象局2010年新技术推广项目。

作者简介: 吴焕萍(1977—),男,博士,高级工程师,主要研究领域为气象信息技术;吕终亮(1981—),男,工程师,主要研究领域为GIS应用;张华平(1978—),男,博士,副教授,主要研究领域为自然语言处理;罗兵(1967—),男,高级工程师,主要研究领域为气象信息技术。E-mail:whp@pku.org.cn

收稿日期: 2012-09-04 **修回日期:** 2012-11-02 **文章编号:** 1002-8331(2014)13-0247-05
CNKI **网络优先出版:** 2012-11-28, http://www.cnki.net/kcms/detail/11.2127.TP.20121128.1453.007.html

1　引言

气象服务产品具有直观、形象、简单易懂的特点,但同时要求精细化、个性化、多样性、时效性、主动性。对于公众来说,气象数据或者相关图表过于专业与复杂,需要领域专家进行解读与提炼,最终形成自然语言表述的气象服务文本信息,也是公众最容易接受的气象服务形式[1]。因此,中国气象局日常发布的国内外“天气公报”、“海洋天气公报”、“重要气候信息专报”、“天气服务公报”等诸多气象服务产品中,均体现了文本语言(或者称“文本”,下文将视为同一概念)描述为主,辅以图形或者表格说明的基本行文原则。天气预报与气象服务产品注重“图文并茂”,但从目前技术发展来看,图形的自动化生成方法在气象信息科学可视化技术发展下相对成熟,如MICAPS和MESIS均具有较强的图形产品自动生成能力[2],而文本的自动化生成方法研究还远远不够。目前业务上依然是预报员人工撰写,甚至看图说话来完成,这种人工编写效率极其低下且常常满足不了时效性要求,还会由于预报员知识背景差异等原因导致文本内容出现偏差。因此,面向气象服务领域内准确、高效、合理、符合自然语言表达的文本生成技术亟待深入研究。

国外于20世纪70年代初就已经开始重视了天气预报文本的计算机自动生成研究。最早的天气预报文本生成器采用了文字替换法(CWF),其代表性的有IFPS、RAREAS、MarWords、Scribe等业务应用系统;随后20世纪90年代初开始引入自然语言处理技术(Natural Language Processing,NLP),一些面向特定天气预报领域的文本生成系统如Forecast Generator(FoG)、SumTime Meteo等也得到了一定发展与应用[3]。相比之下国内相关领域的研究则开展较晚,气象部门主要使用了从简单数据到文本表格形式的预报文本生成,如采用了从天气代码直接到对应文字描述的简单转换。真正意义上基于自然语言处理技术的成果,可以追溯到2000年上海交通大学开展的多语种天气预报文本自动生成系统(MLWFA)的初步研究。总体来看,国内外这些研究为面向气象领域的自然语言文本生成进行了有益探索并奠定了一定的基础[4-8]。

结合气象服务气象区域文本描述的基本业务需求,本文分析了中央气象台每日发布的大量历史“天气公报”文本内容,提出了综合运用自然语言处理和地理信息分析方法来形成文本语言自动生成方法[9-11],并对所涉及的历史文本内容分析与特征提取、地理区域划分、气象要素空间分析、文本组织与生成等关键技术问题进行了深入讨论。通过探索文本类服务材料的计算机自动或者半自动生成方法,生成效率的提高将有望将预报与服务人员从繁重、重复的体力劳动中解脱出来,使其有更多的时间和精力用于真正思考如何更好地做好预报与服务,从而提高预报准确率;另一方面,也将有望进一步拓宽服务材料的生成领域,实现服务材料的多形式表达与快速多渠道发布,不断满足用户精细化和个性化的需求。

2　基本原理

2.1　气象落区描述

气象落区是指某一气象要素在某一地理区域内发生的位置,气象落区文本描述则是指一定的地理区域上所发生的天气气候现象及强度的文字性说明,一般包括对过去发生的气象实况进行总结描述,也包括对未来预报的气象要素发生区域进行展望描述,如中央气象台每日发布的指导预报产品“天气公报”中,其主要内容是对未来三天的降水预报落区及变化趋势进行描述,如图1所示。

图1　中央气象台天气公报部分内容示意图

通常气象落区文本描述的信息源于可以分两大类:一类是实时气象台站观测信息;另一类是气象预报信息。这些信息经过预报员的大脑解译与分析后形成的天气实况或者气象预报文本,不但要求其文本在时间、地区及方位、气象要素种类(如降水、温度、湿度等)以及相应量级(如小雨、中雨、暴雨等)等方面合理、准确并符合自然语言表达,还要求符合预报员长期以来形成的语言表达习惯,因此具有较高的语言特征和用语要求。

2.2　基本原理

气象落区文本描述的计算机自动生成方法,就是要解决从气象数据到文本数据的生成问题。本文提出了以下基本思路:首先将气象观测数据或者预报数据通过一定的模型转化到空间区域上,即一定的气象落区,然后与一定的地理区划数据进行GIS空间分析,确定特定气象要素及相应的量级落在某地理区划上,最后运用自然语言生成技术(NLG),对气象要素的时间、地点、强度等信息进行合理组织,并运用段落规划、句子规划、句子优化以及相关后处理形成自然语言表达的气象落区描述文本。上述流程如图2所示。

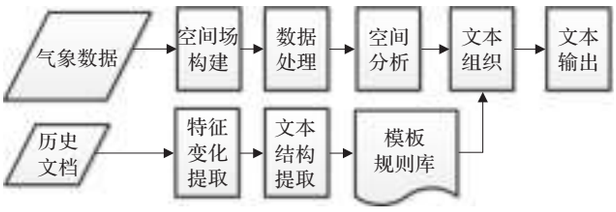


图2 技术流程图

3 关键技术分析

3.1 文本分析与特征提取

文本分析与特征提取是为了分析抽取某一类文本的内容与结构模板,并为最终生成的文本内容提供规则库。对于气象落区文本特征的抽取,一般需要对文本中所涉及的时间变量、地理变量、方向变量、气象要素变量,以及短句结构特征、句子、段落和篇章等元素进行有效特征提取。

本文共对业务人员人工撰写的1963个历史"天气公报"文档进行了自然语言统计学分析,抽取出来了气象要素、地理及方位变量、描述气象的短句与结构特征。其中,气象要素变量包括描述各种气象信息的天气名词及相应的强度(量级),如降水量及"小雨"、"中雨"、"大雨"等。地理变量包括了描述地理位置的地名名词,如华北、内蒙古、长江流域等;方位变量包括了大部、局部、东部、西部、南部、北部、中部、中大部等名词;短句结构特征是指描述气象要素所涉及上述变量的通用的句子表达形式,并同时经过短句结构的抽取形成了句子模板库。图3所示是句子模板库的一个简单例子,中括号(即[])及其中间的信息表示一个变量,变量有[地点]、[级数]、[方向]、[雨雪强度]等。图3中可以看出,对于天气现象风的句子结构,有如什么海域将多少级的风,或者地方有多少级的风,某些地方风力达到多少级,或者什么地方有多少级阵风、方向多少。

Weather = (风) (雨雪) (气温) (能见度)
风 = [地点]{将}有[级数]{左右}[方向]风{,部分海域阵风可达[级数]}。[地点]{将}有[级数]{左右}风{,[地点]的风力可达[级数]}。[地点]有[级数]{,[地点][级数]}[方向]风。[地点]{将}有[级数]、阵风[级数]的[方向]风。
… …
气温 = [地点]气温将[升降][度数]℃{,其中[地区][升降]温幅度可达[度数]℃}。[地点]以[天气]为主,气温逐渐回升。[地点]气温将下降[度数]℃,局地降温幅度可达[度数]℃以上。[地点]将出现[度数]℃以上高温天气,{其中,[地点]最高气温可达[度数]℃。}
能见度 = {[DAY]日早晨和夜间,}{今天早晨,}[地点]{将}有[大气透明度]{,[地点]能见度不足[能见距离]米}。 {今天早晨到上午,}[地点]有能见度不足[能见距离]米的雾,[地点]能见度不足[能见距离]米。{今天早晨到上午,}[地点]有[大气透明度]{,其中,[地点]有能见度小于[能见距离]的雾。}

图3 天气公报句子结构特征

同时还对"天气公报"的结构特征进行了抽取,主要的特征如图4所示。图4中篇章结构表明,通常说明了什么时间(具体到小时)发布的预报,签发的预报员,主要天气原因,预报的时效,天气趋势总结,具体预报内容等主要内容,其中具体的预报内容则结合图3所示的句子结构来组织。

天气公报
中国气象局中央气象台
预报:[AUTHOR]
[YEAR]年[MONTH]月[DAY]日[HOUR]时
[TITLE]

{受[固要]影响,}未来[天数]天,[WEATHER_SUM]。

具体预报如下:
[DAY]日[HOUR]时至[DAY+1]日[HOUR]时,[WEATHER_DET]。
[DAY]日[HOUR]时至[DAY+1]日[HOUR]时,[WEATHER_DET]。
[DAY]日[HOUR]时至[DAY+1]日[HOUR]时,[WEATHER_DET]。

联系方式:Tel:[PHONE] Fax:[FAX]

图4 篇章结构特征

3.2 地理区域划分

地理区域是用于描述某种天气现象所在的空间区域,它的划分直接决定了文本生成的内容是否符合自然语言特征。一般来讲它的划分原则既要结合气象领域的全国气象地理区划标准[12],同时也要考虑预报员多年来形成的语言表达习惯。

全国气象地理区划主要分为四级,其中全国一级气象地理区域有:西北地区、华北地区、内蒙古地区、东北地区、黄淮地区、江淮地区、江南地区、江汉地区、华南地区、西南地区和西藏地区。全国二级气象地理区域是在全国一级气象地理区域基础中按方位进行划分的,如西北西部等。各行政省份或直辖市划分为全国三级气象地理区域,全国四级气象地理区域则在三级气象地理区域的基础上按方位进行划分,如江西南部、中部、北部。理论上来讲,可以直接运用标准的四级气象地理区划来分级表达落区,但通过对天气公报的地名统计分析,发现除使用全国四级气象地理区划中所规定名称外,预报员多年来形成的习惯还常使用地名和河流、平原、山脉和高原等名称描述该地域的气象信息,如青藏高原等。因此,结合对地名的统计分析与识别的结果,运用地理信息技术空间分析方法对上述标准的地理区域进行了合理调整,形成了面向气象落区专用的地理区域划,以进一步符合预报员描述习惯。对于中国海域部分,主要分为:渤海、黄海、东海、台湾海峡、南海、北部湾等海域,则不再细分子二级区域。

3.3 气象要素的空间分析

气象要素的空间分析包括了其本身的空间化和空间化后的气象要素与地理区域的空间叠加分析,它的分析结果决定了文本内容是否准确。

根据中央气象台的业务流程,天气指导预报为落区预报,已经表达了一定的地理空间未来可能的发生的天气信息,而对于气象观测类型的数据,可以采用合适的

客观化分析模型生成基于空间区域的气象分布。对于降水量、温度等连续变化量的客观化,一般可以采用CRESSMAN插值以及IDW(反距离加权平均)等插值方法,而对于雾等离散变化的天气现象量的客观化,一般可以采用泰森多边形法(Thiessen)的方法来确定空间分布。

气象要素与地理区域进行空间分析可以确定气象要素所发生的空间区域。这里主要运用了气象要素空间分布场与多级地理区域进行相交(Intersect)分析、融合(Dissolve)分析等,这样可以获得不同地理区域上每类气象要素的类型、量级、和面积大小等信息。

针对预报员在描述预报文本时尽量采用某地区大部或局部等模糊量词的特点,本文采用"叠加度(P)"及大小来表达大部和局部等概念,即气象要素数据与其覆盖地理区域面积之比。叠加度的引入可以进一步判别是否需要按一定的精度来输出文本。具体空间分析时,采用了首先将气象要素逐一与四级地理迭代空间分析,然后根据"叠加度"判断是否需要进行二级地理区域的空间分析,同理是否采用三级、四级区域进行再次空间分析。该方法一方面加快了效率,还在空间分析阶段就保证了同一区域没有被重复处理。

此外,天气预报未来三天趋势分析时,需要分析气象要素在时间尺度上的空间变化,如降水量未来三天将从东部逐步转移到西部地区。取气象要素空间分布场的内点,然后判断其空间方位以及空间位置的变化,为了处理简单这里只考虑了最大量级的气象要素的空间变化。

3.4 文本组织与生成

自然语言生成(Natural Language Generation,NLG)方法能够从要表达的意思出发选择词语,生成符合语法和逻辑,内容行文流畅,符合人们理解的句子,通常采用了内容规划(Document Planner)、句子规划(Mircoplanner)、表层生成(Surface Realize)的流水线式计算机模型[10]。其中,内容规划主要确定文本的内容,句子规划则主要通过省略、指代、合并等手段使规划的文本更加通顺、自然,表层生成则最终输出文本。

对于"天气公报"的内容规划,本文采用了简单模板方法即通过对历史文本的特征提取来形成了一定的模式与规则;对于句子规划,语句的先后顺序需要遵循以下规律:

(1)地理区域空间的描述顺序。全国范围内的总体方向主要是先由西向东,再由北到南,如一级地理区域主要依次为西北地区、西藏地区、内蒙古地区、东北地区、华北地区、黄淮地区、江淮地区、江汉地区、江南地区、华南地区、西南地区。

(2)地理区域分级的描述顺序。先是全国一级气象地理区域,接着是全国二级气象地理区域,依次类推。但对风要素来看,地理位置包括了陆地与海洋区域,一般顺序为先为大陆然后为海洋。

(3)气象要素类型的描述顺序。主要顺序为降雨、大风、降温以及其他天气现象。

对于气象要素的数值量级的描述,结合气象部门的业务规定也逐一转换成文本词语,如降水量不同的数值范围替换成小雨、中雨、大雨、暴雨等词语。

对最终输出的语句需要进行合并等后续优化处理,这里主要涉及了同一地理区域、相同的量级大小等语言合并规则。通过对比分析大量历史天气预报图形和其相应的描述文本,设计区域合并规则如表1所示,并引入输出"大部"、"局部"等词语来模糊描述地理区域[13]。值得注意的是,合并时还一并考虑了地理区域本身的空间包含关系,使得文本表达更为合理。

表1 区域输出规则

叠加度(P)	$1.0<P<0.9$	$0.7<P<0.9$	$0.05<P<0.7$	$P<0.05$
一级区域	该区域	大部	转下级区域	—
二级区域	该区域	大部	转下级区域	—
三级区域	该区域	大部	转下级区域	—
四级区域	该区域	大部	局部	—

4 技术实现

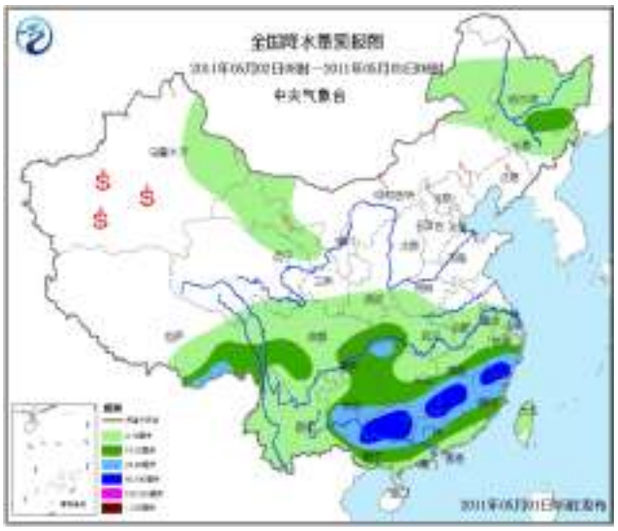
气象落区文本特征提取是一种典型的文本信息抽取(Information Extract),如时间描述(早上、中午、晚上、昨天、明天等),空间描述(区域、省级、市级等),方位描述(西北、东南、南部、大部、局部等),天气要素描述(降水、温度、风)等。本文采用了开源的ICTCLAS分词工具(它包括了中文分词、词性标注、命名实体识别、新词识别等主要功能)进行地名提取、气象变量提取等[14]。同时,进一步开发了词频统计等工具进行语句结构与模式的提取以形成模板。

文本生成过程需要大量的空间数据处理操作,GIS发挥了重要作用。本文采用了开源空间分析包GEOS[15](Geometry Engine,Open Source),它是对OGC规范中简单几何要素对象操作的C++语言实现,是一个集合形状的拓扑关系操作实用库,主要实现了空间关系(相等、相交、包含)和空间叠加分析(缓冲区、交叉分析)操作等,能够较好地实现本文空间分析。具体分析时,将所有气象数据(MICAPS14类交预报员格式的数据))以及地理区域数据转换成GIS格式,然后直接应用该空间分析引擎即可。空间分析的结果采用XML格式进行定义与保存,并最终参与文本生成。

天气公报文本生成采用了基于规则的文本生成思路,将气象信息空间分析化的结果与文本结构模板进行关联并形成较好的自然语言表达的文本内容。

结合上述文本生成方法,采用C++语言对"天气公报"的气象落区文本的生成进行全部编程实现,并具备自动定

时运行能力，程序输入数据主要包括了中央气象台每天的未来三天降水预报、灾害性天气预报落区的等业务数据。



2日08时至3日08时，西北中东部、新疆东部和北部、西藏东部和南部、内蒙古局部、东北中北部、黄淮西部、江淮大部、江汉、江南、华南和西南等地有小到中雨，其中，西藏东部、湖北南部和西部、湖南局部、江西中南部、浙江中南部、福建部分地区、广东局部、广西大部、重庆局部、贵州局部和云南东南部和东北部等地有大雨，其中，湖南南部、江西中南部、浙江南部、福建西北部和东北部、广东西北部和广西部分地区等地有暴雨。

2日08时至3日08时，新疆东部、青藏高原东部偏南地区和东北部、内蒙古东部和西部、东北地区中北部、甘肃中西部和东南部、陕西南部、黄淮西部、江汉、西南地区、江淮大部及其以南大部地区有小到中雨或阵雨，其中，江汉西南部、贵州南部、江南南部、华南北部、西藏东南部等地的部分地区有大雨，局地有暴雨。

图5　计算机与人工撰写对照

（上为预报图，中间为人工撰写，下为计算机生成）

5　问题讨论

自然语言生成领域通常采用正确率（如生成系统是否表达输入的全部意思）、通顺度（如生成的文本是否通顺，文法是否正确，文章风格是否符合用户要求等）、任务评估（生成系统应用于实际领域中的代价、社会影响等）等指标试图来评价生成系统的质量。由于通用文本生成方法本身还存在较大的技术难点，因而其相应的定性量化评估方法更是远远不成熟，上述指标也仍然停留在定性化评估研究阶段[16-17]。本文借鉴正确率和通顺度两个方面评价内容，对比分析了2011年4月—2011年9月以来由预报员和计算机分别生成的文本内容，总体可以看出：

（1）自动生成的文本内容正确，在落区描述方面甚至比预报员人工撰写的预报文本更加描述细致（其精细化程度由本文提出的“叠加度” P 取值决定），主要体现在不遗漏重要的气象要素所在的地理区域及相应的量级，相比之下，预报员在撰写公报时则主要考虑总体趋势表达从而做到行文简洁。这一点上又不太符合天气预报“模糊语言”原则的文本描述习惯[13]，因此自动生成的文本某种程度上还显得“冗长”，尤其是那些复杂的天气形势。

（2）自动生成的文本内容总体符合了预报员习惯，如空间区域的分级描述和空间区域先后顺序的描述，文本内容也较为通顺、语义、语法正确，文本风格也符合了预报员行文习惯。

（3）自动生成的文本内容固定、形式单一（由模板和规则库决定），而预报员人工撰写内容时还经常结合预报经验和领域知识做相应补充，如落区量级表达时常对其局部地区进行补充说明。例如2011年8月29日天气公报中有“其中，浙江东南部、福建东部、台湾等地的部分地区有暴雨，局部大暴雨，雨量一般有80～150 mm，台湾南部局部雨量可达200～400 mm；上述部分地区并伴有短时雷雨大风等强对流天气”等补充性描述（下画线部分），而这些信息仅仅依靠现有的输入信息自动生成是远远做不到的。

上述结论也得到了负责撰写“天气公报”的中央气象台天气预报室短期科等多位同事认可，并总体认为生成效率高，具有一定的参考性和实用性，可以作为天气公报中文本材料的初稿。图5为2011年5月1日的降水量预报落区的文本对比分析示意图。2011年5月1日属于气象业务中汛期气象服务时段，因此气象落区从降水量量级、空间分布、范围等来看均具有一定的代表性和复杂性。

6　结束语

本文紧紧围绕气象落区文本生成系统的主要问题，即哪些内容应该包括在生成系统的输出里，以满足预报员的撰写意图，如何保证生成内容的连贯性，如何保证生成内容在语法和语义上的正确性等；结合问题提出了气象落区文本生成方法与流程，并重点阐述了文本分析与特征提取、地理区域划分、气象要素空间分析、文本组织与生成等四方面关键问题，同时还给出了相应的技术实现和初步的评价。总体来看，计算机自动生成的天气预报落区文本虽然还不能与人工撰写的内容“媲美”，但可以作为预报员人工撰写文本的初稿，预报员在此基础上再作进一步的润色修改即可以成为最终对外服务的指导产品。下一步还将深入开展应用分析与评估，并研究将现有的文本生成功能集成于MICAPS、MESIS或者CIPAS（气候信息交互显示与分析系统）等业务系统中，形成文本辅助生成工具（如生成天气预报文本、实况文本等信息）供业务用户使用。

本文综合运用了自然语言处理、地理信息科学等交叉方法对计算机自动生成气象落区的文本进行了初步的有益探索，其面向特定领域的计算机文本生成方法具有一定通用性，这也为进一步深入探索面向气象服务领域的文本生成开辟了新思路和研究方向。

（下转266页）