**基于地理信息系统的气象文本生成与交互研究**

**李笑侃**

**2012年12月**

**基于地理信息系统的气象文本生成与交互研究**

作 者 姓 名 李 笑 侃

学 院 名 称 计算机科学技术院

指 导 教 师 李 玉 岗

答辩委员会主席

申 请 学 位 工学硕士

学 科 专 业 计算机科学与技术

学位授予单位 　 北京理工大学

论文答辩日期 2012年12月

**The Study of Meteorological Text Generation and Interaction Based on Geographic Information System**

Candidate Name: Xiaokan Li

School or Department: Computer Science and Technology

Faculty Mentor: Yugang Li

Chair, Thesis Committee:

Degree Applied: Master of Engineering

Major: Computer Science and Technology

Degree by: Beijing Institute of Technology

The Date of Defense: December, 2012

**基于地理信息系统的气象文本生成与交互研究 北京理工大学**

**研究成果声明**

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签 名： 日期：

**关于学位论文使用权的说明**

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的,复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签 名： 日期：

导师签名： 日期：

**摘要**

对于公众而言，气象原始数据、图表过于专业与复杂，迫切需要领域专家进行解读和提炼。气象服务文本注重“图文并茂”，但是文本的自动、高效、合理、准确的生成方法离实际应用需要相距甚远。本文立足于自然语言自动化生成的算法研究，将自然语言生成与气象领域结合，旨在将复杂多维的地理数据提炼成特定格式的气象数据，生成浅显易懂的气象服务文本信息，并进一步研究探索受限领域的语言特征分析方法与语言生成算法。

本文首先对“重要气候信息”、“天气公报”、“每日天气提示”这三种气象服务文本进行词和短语的特征提取，分析了这三种文本的用语，提取了预报员行文的地理要素和气象要素特征词，讨论了这些要素对业务人员撰写行文的影响，并建立了气象公报的语言模型；其后，本文采用空间分析技术，对地理数据进行分析处理，生成特定格式的气象数据；本文最后以国家气象中心气候交互显示与分析系统为依托，将空间分析技术与文本生成相结合，通过分析任意气候落区的气象观测站的实时观测数据，生成了与之相对应的气象服务文本。

本文提取了气象服务文本的用语特征，解决了通用自然语言生成与气象领域相结合的难题，为气象文本的自然语言表达与计算自动生成奠定了良好的技术基础；系统实现了地理数据与气象数据之间的自动化转换、分析，已为气象业务中心提供科学决策依据，体现贯彻了以气象科技为支撑的战略发展方针。

**关键字：**地理信息系统，空间分析，文本生成

**Abstract**

For the public, the meteorological raw data and charts are too specialized and complex. These problems urge the necessity for experts in this field to interpret and to refine. The meteorological service mainly focus on the “illustrated” text, but actually, the method of textual generation which include automation, efficiency, reasonability and accuracy had far exceeded the actual application needs. This paper is based on algorithm which automatically generated by natural language, and combines both the generation of natural language and meteorological fields. It also aims at refining the complex geographic data with many locations into meteorological data with a specific format, so the complex data can be turned into the meteorological service text information which is easy to understand. Further, it also studies the linguistic features analysis methods and language generation algorithm in some restricted areas which cannot be explored.

Firstly, this paper gives a characteristic extraction of words and phrases of three meteorological service texts named “important climate information”, “weather bulletins”, and “daily weather prompts”. And then it analyzes the terms of the three texts and extracts some geographic and meteorological characteristics of words of the forecasters, establishes the meteorological bulletins language model. Subsequently, this study wields spatial analysis technology and gives analysis of geographic data to generate meteorological data with a specific format. Finally, the article relies on climate interactive display and analysis system of the National Meteorological Center and combines spatial analysis techniques and text generation and generates meteorological services correspondingly through the analysis of real-time observation data of meteorological stations of any climates.

This study extracts text language characteristics of meteorological services, solves the problem of combination of general natural language generation and meteorological fields, and builds a good technical foundation for weather text natural language expression and computing automatic generation. It also achieves the automatic conversion and analysis between geographic data and meteorological data, and provides a scientific basis of decision making for meteorological operations center, embodies the implementation of strategic approach which is supported by meteorological science and technology.

**Keywords:** geographic information systems, spatial analysis, text generation

**目录**

[第一章 绪论 1](#_Toc343094454)

[1.1 选题的研究意义 1](#_Toc343094455)

[1.2 国内外研究状况 4](#_Toc343094456)

[1.3 本文的主要研究内容 6](#_Toc343094457)

[第二章 气象服务语言的特征分析 8](#_Toc343094458)

[2.1 气象服务文本特征抽取方法 8](#_Toc343094459)

[2.1.1 语料库选取 8](#_Toc343094460)

[2.1.2 特征提取 9](#_Toc343094461)

[2.1.3 特征提取实验步骤 12](#_Toc343094462)

[2.2 特征与用语分析 14](#_Toc343094463)

[2.2.1 地理区划特征 15](#_Toc343094464)

[2.2.2 地理方位和范围特征 16](#_Toc343094465)

[2.2.3 气象要素特征 17](#_Toc343094466)

[2.3 三类服务材料用语分析 19](#_Toc343094467)

[第三章 GIS空间分析与气象原始数据生成 20](#_Toc343094468)

[3.1 GIS空间分析概述 20](#_Toc343094469)

[3.1.1 地理信息系统 20](#_Toc343094470)

[3.1.2 地理空间数据库 21](#_Toc343094471)

[3.2 气象数据接口规范 22](#_Toc343094472)

[3.3 气象服务文本的空间推理分析 26](#_Toc343094473)

[3.3.1 气象地理要素 26](#_Toc343094474)

[3.3.2 气象地理区划 27](#_Toc343094475)

[3.4 气象原始数据生成 30](#_Toc343094476)

[3.4.1 地理原始数据 30](#_Toc343094477)

[3.4.2 地理数据分析与处理 31](#_Toc343094478)

[第四章 气象公报文本生成 36](#_Toc343094479)

[4.1 气象公报结构分析 36](#_Toc343094480)

[4.2 跨领域的语言自动化生成模型 36](#_Toc343094481)

[4.3 受限领域的文本自动化生成 37](#_Toc343094482)

[4.3.1 短语规划 38](#_Toc343094483)

[4.3.2 句子规划 38](#_Toc343094484)

[4.3.3 篇章规划 42](#_Toc343094485)

[4.4 图像与文本交互 42](#_Toc343094486)

[第五章 原型系统与实验验证 44](#_Toc343094487)

[第六章 总结与展望 47](#_Toc343094488)

[参考文献 48](#_Toc343094489)

[攻读学位期间发表论文与研究成果清单 51](#_Toc343094490)

[致谢 52](#_Toc343094491)

# 第一章 绪论

## 1.1 选题的研究意义

随着社会与经济的发展，人们对气象关注越来越多，对气象服务产品的要求也越来越高，同时国家对气象科技发展也给予了前所未有的高度重视，气象业务在全社会各个领域发挥着越来越重要的作用，与社会和谐进步和生态环境可持续发展密切相关。气候的预报预测是基础能力，是所有气象业务产品向社会提供服务的出口，是气象工作的出发点和归宿[1]。

2010的甘肃舟曲特大泥石流灾害、西南特大干旱、2011年6月长江中下游地区旱涝急转等重大气象灾害凸显了气象服务重要性与迫切性。气象部门这种以气象服务为需求引领的工作正在全面展开，在全面提高预报准确率的基础上，推动公共气象服务的发展，尤其是推进气象服务产品的个性化、精细化、时效性、多样性、主动性。公共气象服务材料具有直观、形象、简单生动的特点，而且又与专业的预报分析材料不同，主要原因在于它一般不进行复杂、深入的天气系统、天气形势等分析与描述[2-3]。当今公众气象、决策气象、专业气象以及科技气象等重要公共气象大多都采用 “图文并茂”的表现形式，主要有如下特点。

1) 图像形式的气象服务生动、直观，成为了表达气象服务信息的一种重要方式。

对于大多数用户来说，气象数据、图表过于专业与复杂，迫切需要领域专家进行解读、提炼，最终形成自然语言表述的、浅显移动的气象服务文本信息。如图1-1所示，尽管图(a)中清晰的标出每个站点降水量值，而图(b)进一步给出了降水分布图，但由于缺少文本方面详细的说明和描述，用户依然很难迅速判断哪个站点的降水量最大以及站点名称，以及暴雨量级空间区域的位置。

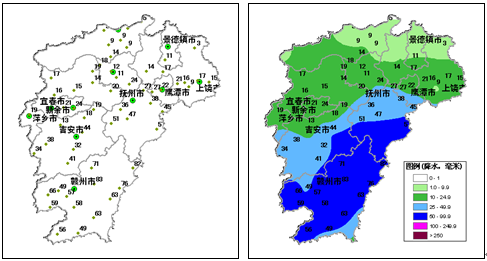


图 1-1 降水分布图

还有研究表明[4]，对于大多用户而言，人脑能获取最精确的解释信息来自于气象文本描述内容，文本描述具有信息量大而准确、内容简短而传输快捷的重要特点，因而，自然语言表达的文字表现形式，一直是面向气象服务的重要手段。中央气象台等单位日常发布的“天气公报”、“重要气候信息”、“每日天气提示”等诸多气象服务产品中，均体现了汉语本文描述为主，辅以图形、表格的基本行文原则。

2)从气象部门的业务要求来看，气象产品千变万化，几乎没有相同的气象报告，撰写准确的合理的气象服务文本需要较强的气象专业知识以及实际工作经验，而且存在大量机械重复的工作。

根据业务流程，气象部门每天实时获取大量的各类气象观测资料与模式预报资料。这些观测资料经过预报员的综合分析，产生服务于不同公众的气象信息。针对不同的服务对象，业务人员开始编写各类预报服务材料。对于图形图像产品，可以从各类信息系统的输出系统快速得到，但对于文本内容，却依然需要依靠预报人员在较短的时间内通过人工撰写来完成。

通常服务文本需要对大量的历史天气以及实况天气进行总结和描述，不仅需要“看图说话”来形成文本内容，还需要对气象信息进行充分的描述甚至初步的翻译（尽管不需要复杂的天气形势分析）。故撰写准确合理的文本材料对预报员来说是复杂而又机械的劳动，主要存在如下问题：第一，撰写效率低下，满足不了日益增长的气象服务的需求，尤其是时效性方面；第二，气象观测资料是海量的，预报员之间的行文风格、素质水平也有差异，同时处理时间比较紧迫，这些最终都可能导致撰写的文本丢失关键性信息或者服务文本的语言风格迥异，从而导致服务质量的偏差。

3)文本的自动化、高效、合理、准确的生成方法离实际应用需求相距甚远。

随着计算机技术特别是图形图像技术的发展，气象信息科学可视化技术也得到飞速发展，因而图形图像形式的服务产品的自动或者半自动加工不再是主要技术问题。“图文并茂”的“图”技术得到了较好的发展与应用，气象部门广泛应用的气象信息综合分析处理系统(Meteorological Information Comprehensive Analysis and Process System, MICAPS)和近年组织开发的气象服务系统(Meteorological Service Information System, MESIS)均能够加工制作高质量的图形图像产品。然而，面对气象服务内容的多样性（如：气象信息既涵盖了实时与历史的观测信息、还有预报预警信息；既有单点的站点信息，又有客观数值预报的格点场分布信息，还有预报员主管的落区预报），气象服务产品中的“文”的自动或者半自动生成技术由于长期缺乏有效深入的研究与探索，却远远没有得到较好的发展与应用。

4）通用自然语言生成技术多年进展一直不大，但自然语言生成技术与受限领域结合，特定领域范围的限定规避了通用语言的泛化与歧义问题。

自然语言处理(Natural Language Processing, NLP)，近年来得到了较好发展，但作为通用的自然语言生成技术研究多年来一直没有太大的进展。但是将受限领域与通用自然语言领域相结合，可以取得良好的研究成果，为气象文本生成与交互提供了明确的研究方向。

综上，气象服务文本生成方法研究的价值与意义，已经成为众多领域专家和部门领导的共识，也体现、贯彻了以气象科技为支撑的战略发展方针。如2010年国家气象局长在工作会议上进一步指出[5]，“切实提高公共气象服务质量，抓好公共气象服务业务系统建设，组织重大科技攻关，推荐公共气象服务产品系统建设，不断满足用户精细化和个性需求。”同时，2010年国务院出台的《国家气象灾害防御规划》中也明确指出“完善气象灾害信息网络，建立和充分利用各类现代化媒体、通信工具和资源进行气象灾害预警信息的发布机制和平台”。从长远看来，研究探索文本类服务材料的自然语言生成与交互方法，将进一步的拓宽服务数据的生成领域，提高服务材料生成效率，实现服务形式的多表达与及时多渠道发布，最终将有效提升公共气象服务能力。

基于上述内容，本文拟建立气象服务文本生成与图像交互系统，旨在将复杂多维的地理数据与气象数据进行精准的匹配，并解决通用自然语言生成与气象领域相结合的难题，为气象业务发展提供科学决策依据。

## 1**.2 国内外研究状况**

本文的研究内容主要包含两个方面，一是GIS空间分析，二是受限领域的自然语言生成。GIS空间分析技术已经发展的相对成熟，而受限领域的自然语言生成却进展缓慢，本节主要介绍自然语言生成的技术以及气象公报自动化生成技术的发展历程。

上世纪60年代开始的自然语言处理技术[6-8]，已经成为了当前的研究热点，成为了一种重要的文本生成方法。NLP属于人工智能和计算语言学分支，其处理的核心问题是语言的自动理解(Language Understanding, NLU)和自动生成(Language Generation, NLG)。NLG语言生成系统采用了内容规划(Document Planner)、句子规划(Sentence planning)、表层生成(Surface Realize)的流水线计算机模型（也可称为管道模型，Pipe Planner）[9-10]。其中，内容规划主要确定文本内容，往往有两种方法：Mckeown的Schema方法和Mann、Thompson的修辞结构理论即Rhetorical Structure Theory。句子规划主要通过省略、指代、合并等手段使规划的文本更加自然通顺。NLG的基本思路是从抽象的概念层次开始，根据应用目标和用户模式，通过选择并执行语义和语法规则来生成文本。

正是由于自然语言生成技术，才促进了人们对气象文本自动化生成的研究。在上世纪70年代，国外开始了对此方面的研究，最早采用的是文字替换（Template-based Computer Worded Forecasting, CWF）的方法，首先要对生成的气象文本信息建立模版，它根据满足用户需要的一般格式来建立，与气象信息有关的文字设置成为模版的槽。在自动化生成文本时，将实际的气象信息转换成适合的文字并填入模板槽内，当所有的模版槽被填满后，气象文本就自动生成了。CWF虽然简单、容易实现，多适用于生成表格文档，但由于需要大量模版，因而管理困难，也不能进行高效的数据分析，更不能形成像自然语言那样流畅的文本，应用范围有限。关于这方面的代表性成果有：IFPS（Weather forecast report generator, 1970）、MarWords(Marine forecast report generator, 1988)、1993年的ICWF以及1999年的Siren系统[11]。RAREAS、MarWords等后来发展成了FoG(Forecast Generator)。FoG主要用于生成海洋天气预报文本，基于预报分析平台(FPA)开发，从FPA的数据库抽取数据，经过简单分析，然后将被选择内容发到语言处理中心进行结构化处理，最终形成语言预报文本(English and French)[12-13]。英国阿伯丁大学研究了海洋天气预报和数值气象预报的计算机自动生成技术，采用了、工海洋气象预报文本资料以及相应的气象预报数据，建立气象预报语料库，然后通过分析数据、规划文档等来实现预报文本的自动生成。在此基础上，开发了天气预报应用项目SumTime\_Meteo[14]。SumTime\_Meteo形成预报文本初稿，供预报与服务人员做后编辑，最终形成预报产品。这些成果后来用于与英国国际航空与海洋局（AMI）的合作项目RoadSafe[15-16]，在详尽的天气信息以及复杂的计算机数值模拟的道路表面温度预报信息的基础上，自动生成了冬季道路维护（如播撒盐粒来除道路结冰）的交通路径规划的建议文本。MeteoCogent是一个可配置的一个天气预报生成工具，将天气预报数据转换成英语文本预报[17]，它从模式输出统计预报（MOS）或者一些使用“明确天气要素”的解释性数据可以作为其输入数据，结合用户给定的明确的标准，输出结果可以为编码形式的预报或者语言文本，报告的形式与内容可以扩展定制，允许多种形式的文本输出。

虽然国外较早的开展了气象文本半自动或者自动生成的技术研究，国内关于这方面的研究却依然很少。我国的气象服务（早期称为“有偿服务”）是从上世纪70年代末开始的，而且仅有北京、上海、天津、吉林等几个省市开展的有限专业气象服务。1985年，在国务院转批“关于气象部门应大力开展专业有偿服务和综合经营的报告”之后，气象部门开始大力发展专业有偿气象服务，但是仅限于气象预报中的长、中、短期节目，而且只有文字报告、电台广播、电话等服务手段。在1992年国家确立把气象科技服务作为气象事业的重中之重后，气象服务现代化进程才有了明显的加快。

国内的一些气象部门使用了简单数据到文本表格来形成预报文本 [18-19]。吉林气象科学研究所的全国城市天气预报文本转换系统以及北京城市气象研究所张京江等人开发的表格形式的预报产品，基本上都是天气代码到对应文本信息的简单转换。从国内来看，气象领域内主要的研究成果是1999年由上海交通大学研究的多语种天气预报文本自动生成系统（MLWFA）[20-22]，MLWFA采用气象预报数据为开端，选择文本后根据不同语种特点确定相应的文本以及句子结构，再按照句法结构输出，可以实现汉、英、德三种语言的文本生成。

总体来看，国外基于自然语言生成的、面向气象预报的文本生成方法已经得到了不错的研究成果，而国内气象服务文本生成与交互方法的关键技术研究还没有得到足够的重视。把复杂多维的地理信息数据与气象数据之间进行自动化转换、分析；从大规模气象文本库中提取出气象服务特征用语，解决了通用自然语言生成与气象领域相结合的难题；高效、合理、准确的生成“图文并茂”的气象服务文本等是需要进一步深入研究的问题。

## 1**.3 本文的主要研究内容**

本论文的主要研究内容，将紧紧围绕以下研究目标：

1. 如何提取以及提取哪些特征来表达气象服务文本内容。以往的研究对象主要是英语和法语，针对汉语的气象文本研究尚未成体系。由于文化、语言差距很大，故移植性很差。所以从大规模历史气象文本库中提取气象服务特征用语是解决通用自然语言生成与气象领域相结合难题的基础，也是本文工作的重点。
2. 如何把实时的、专业的、复杂多维的地理信息数据与气象数据之间进行自动化转换、分析。气象服务文本由各类海量信息产生，这些信息既包括空间地理信息，又包括大量与空间密不可分的气象属性信息。而且气象数据从本质上来说也是地理信息，因为气象中的风速、温度、气压等都是相对于具体空间位置而言的，没有地理位置的气象要素是没有任何意义的。因此把复杂多维的地理数据与气象数据之间进行转换分析是文本生成与交互的基础。
3. 如何解决受限领域与通用自然语言相结合的问题。在提取气象服务文本用语特征的基础上，将复杂地理数据转换、分析与文本生成的框架模型匹配，形成气象服务文本生成与交互模型，并进一步实现受限领域语言自动化生成是本文的最终目标。

针对上述研究目标，本文首先对“重要气候信息”、“天气公报”、“每日天气提示”这三种气象服务文本进行词和短语的特征提取，分析了这三种文本的用语，提取了预报员行文的地理要素和气象要素特征词，讨论了这些要素对业务人员撰写行文的影响；其次，本文介绍所采用的GIS空间分析技术，对任意气候落区的气象观测站的实时观测数据进行分析与匹配，并生成标准的气象数据；最后，本文从气象数据中提取气象特征信息，并在气象语言特征库以及规则库的基础上生成符合公众阅读习惯的气象文本，并实现图像与文字的交互。本文的各章内容概要如下：

第一章为概述部分，主要介绍了本文的研究意义，研究目标，国内外的研究现状以及本文的主要研究内容。

第二章对大量历史气象服务文本进行词和短语的特征提取，提取了预报员行文的地理要素和气象要素特征词，讨论了这些要素对业务人员撰写行文的影响，并建立了气象公报的语言模型。

第三章介绍了气象服务文本空间分析推理的方法，定义了气象原始数据的接口规范，以及GIS数据与气象原始数据转换分析的详细步骤。

第四章介绍了受限领域（气象领域）与通用自然语言领域相结合的文本生成方法，并建立了相应的自动化生成的语言模型。

第五章介绍了基于GIS的气象文本生成交互系统，并对文章提出的研究目标给予实验验证。

第六章对本文的工作进行了总结和展望。

# 第二章 气象服务语言的特征分析

气象文本信息是气象专家根据不同公众的需求，对大量历史天气、实况天气进行分析和总结，最终提炼成准确易懂的、由自然语言表述的服务文本，是公众最容易接受的气象服务形式。对于大多用户而言，人脑能获取最精确的解释信息来自于气象文本描述内容，文本描述具有信息量大而准确、内容简短而传输快捷等重要特点。如天气气候部门的业务产品“天气预报信息”、“重要气候信息”和“每日天气提示”等，均体现了文本语言描述为主，辅以图形或者表格说明的基本行文原则。气象服务文本面对的人群十分广泛，对公众的影响特别显著，是气象服务业务发展水平的一个体现。气象部门的业务人员对气象观测资料、模式预报资料等进行综合分析与描述，进而产生面向不同服务对象的气象服务文本信息。目前大量的文本基本依靠人工撰写，导致效率低下，还远远满足不了业务需求。

针对上述问题，本章运用自然语言统计学研究方法，选取气象部门（天气和气候领域）人工撰写的典型文本作为研究对象，对大量历史的气象文本中的词和短语进行了特征抽取和综合分析。本章对气象文本中地理区划、方向、范围、气象要素等用户特征的提取和分析研究成果，为气象文本的自动化生成奠定了技术基础，另一方面其文本特征也将对业务人员尤其是新预报员的人工撰写有较大的指导意义。

## 2**.1 气象服务文本特征抽取方法**

在实际的气象服务业务开展过程中，气象服务文本材料越来越多，如何从大规模气象文本库中提取出气象服务用语特征是气象预测、预报业务中的一个值得研究的难题。本章节将主要介绍本文所采取的文本特征抽取的方法及步骤。

### 2.1.1 语料库选取

语料库就是一个由大量在真实情况下使用的语言信息经过科学的收集和组织而集成的专供研究使用的语言资料库。语料库并非语篇的简单堆砌或集合，它应具有以下几个基本特征：（1）样本代表性，（2）规模有限性，（3）机读形式化。语料库有不同的加工层次，加工的语料库一般是指标有语言学标记的语料库。语料库的选择与分类将直接影响文本特征提取的结果。

如表2-1所示，本文选取的语料库包括三种典型的文本材料：2008-2009年间由国家气象局发布的所有天气公报信息（天气预报产品）、2008-2011年间由国家气象局发布的所有重要气候信息（气候服务产品）、2008-2011年间所有每日天气提示（公众服务）和一年内国家发布的所有的人民日报四个部分。

表2-1 语料库类别和用语来源

|  |  |  |  |
| --- | --- | --- | --- |
| 语料库来源 | 语料库类别 | 主要内容 | 用户 |
| 天气公报信息 | 气象服务文本语料库 | 指导产品，预报未来三天内的国内天气情况，主要包括降水、气温等要素的预报 | 指导产品，面向全国、省级业务人员等 |
| 重要气候信息 | 结合当前热点，总结近期国内重要气候服务，每一期一般以一个服务主题为主，如“干旱” | 主要面向决策服务人员及部门领导等 |
| 每日天气提示 | 将每天重要的天气情况描述成通俗易懂的文本形式 | 公众、媒体等 |
| 人民日报 | 普通话用语通用语料库 | 新闻报道、评论、散文、广告等信息 | 公众、机、媒体等 |

### 2.1.2 特征提取

气象服务语言的特征应该具有以下特点：1）特征是能够对文本进行充分表示的语言单位；2）文本在特征空间中的分布具有较为明显的统计规律；3）文本映射到相应特征空间[23]的计算复杂度不太大。气象语言特征提取的研究内容主要包括两个方面：粒度选取与特征提取方法的选择。

**粒度选取：**中文的文本表示通常选用字、词、短语、句子以及句群等更高层次的单位。在实际应用中，到底选择哪种单位来表示文本必须由处理速度、精度要求、存储空间等方面的具体要求来决定。选择的单位越具有代表性，语言层次越高，它所包含的信息也就越丰富，但同时进行分析所付出的代价也就越大。

1）字特征：选择字作为文本特征是最为简单的方法，而且文本分解为字非常容易实现。GB2312规定的汉字共有6768个，因此字特征集合肯定小于6768。字特征的优势在于：相对于其他粒度的特征，它最明显的优势就是集合非常小。当然，它也有非常明显的缺点，从理论上说，一个语意范畴不能由字完整的表示出来，因此字特征对文档的表示能力存在一定的缺陷，有时候会出现严重的偏差。

2）词特征：大多数的特征提取都以词为选择的粒度。原因如下：（1）以词为单位比较符合自然思维习惯，便于利用语言学知识；（2）以词为单位的全文检索方法可以借用英文系统中成熟的理论及方法。词作为特征还有如下优势：词蕴含更为丰富的语义信息，表达文本信息更为完整、准确。而它的缺陷在于，在使用词作为文本特征时，要进行有效的分词和特征抽取，信息处理的工作量和复杂度将明显增大。

3）短语特征：文本表示时，中频词语对于不同主题类文本的区分度都高于高频词和低频词。因此可以使用短语来代替高频词，以降低其出现频率，以词汇类来代替低频词，以提高其出现频率。用短语代替词来表示文本还可以提高对文本的表达能力[24]。组块是介于词和句子之间的语言特征，它是一个类似于短语的句法概念，但从语义上来说，短语要比组块更加完整。Abney[25]和李素建[26]等分别对中英文组块进行了较为深入的介绍和分析，并提出了组块划分、识别和计算的有效方法。

4）概念特征：概念特征是指通过概念标注的方法把有同义和近义关系或上下位关系的词合并为一个语义类，称为概念。虽然使用概念特征描述文本的方法比单纯的词汇能更准确、完整的反映文本内容，但在提取过程中处理起来要复杂得多。

针对大规模历史气象服务信息的文本特点以及规模，本文在特征提取时选择词特征和短语特征混合的方法。既兼顾了词语特征能够完整、准确地表达文本信息的优点，又依靠短语代替词的方法来减少高频词和低频词对文本的区分度低于中频词的问题。

**特征提取方法：**为了提高气象文本信息过滤准确率，在选取词特征和短语特征之后，还需要对特征进行选择，以达到降维和减少计算量的效果。

一个普通的文本能够提取出来的特征词就有几千上万个，因此它的特征空间的维数有几千甚至几万，但大多数的学习算法很难处理这么大的维数。对只有1200篇文章的数据集进行分词后就得到了70000多个特征项，实际应用中的气象服务信息文本集包含的文章确远远不止这么多，所以特征选择的方法的好坏直接关系到研究结果的合理与否。人工选择肯定是最好的特征选择方法，通过这种方法选出来的特征最符合用户的要求，容易获得令人满意的过滤效果。但是，领域专家的选择尺度难以掌握，选取的特征很可能主观性较强，而且研究选择的文档规模特别庞大，手工选择特征变成一件不可能完成的工作。因此，一般采用自动的特征选择方法。

不同的特征对于文档的重要性以及区分度是不同的，高频特征经常会在多个类中出现且分布均匀，故而区分度较小；而低频特征由于对文档向量的贡献较低，因此重要性不高。提高过滤准确率需要去除区分度较小的高频特征，为了加快运行速度则得去除重要性较低的低频特征。研究人员们建立了许多特征评价机制，对特征进行选择。

传统的特征选择的方法很多，其中最简单的是通过词频(Term Frequency)和文档频次（DF: Document Frequency）选择特征。当然，特征评价函数也是经常被研究人员重点使用的特征选择手段的方法。根据某个评价函数计算各个特征的评分，然后按评分值排序，选取评分最高的若干个特征。常用的特征评价函数有互信息(MI: Mutual Information)、交叉熵(Cross Entropy)、信息增益(IG: Information Gain)、统计量(CHI)等。详细介绍如下：

词频方法是将词频小于某一阈值的词删除，从而降低特征空间的维数。词频分析的目的是得到气象服务文本的词语总数以及每个词语的出现次数和频率，进而进行特征分析。

文档频次方法：与词频类似，文档频次是指文档集合中出现某个特征的文档数目。在特征选择中，计算每个特征在训练集合中出现的频次，根据预先设定的阈值去除那些文档频次特别低和特别高的特征。文档频次的计算复杂度较低，能够适用于任何语料，因此是特征降维的常用方法。

交叉熵方法：交叉熵与信息量的定义近似，其公式(2-1)如下：

 (2-1)

交叉熵（KL距离），它反映了文本主题类的概率分布和在出现了某特定词汇的条件下文本主题类的概率分布之间的距离，词汇w的G(w)越大，对文本主题类分布的影响也越大。

互信息方法：互信息是计算语言学模型分析的常用方法，它度量两个对象之间的相关性。假设跟K个主题相关的文档集合分别为C1，C2，…，CK，则特征w对于主题相关Ci的互信息量MI(w,Ci)的计算公式(2-2)如下：

 (2-2)

特征w的互信息越大，说明它与该主题的共现概率越大。对于具有相同条件概率P(w|Ci)的特征，低频特征的P(w)较小，因此互信息比中频特征要高，因此出现频率差异很大的特征的互信息大小不具有可比性。

信息增益方法：机器学习中常常用到信息增益的方法，在过滤问题中用于度量已知一个特征对于主题相关文本预测信息的能力。通过计算信息增益可以得到那些在正例样本中出现频率高而在反例样本中出现频率低的特征，以及那些在反例样本中出现频率高而在正例样本中出现频率低的特征。信息增益G(w)的计算公式(2-3)如下，其中P(w)是词w出现的概率，P(Ci)是第i个目录的概率。

 (2-3)

统计量方法：统计量用于气象度量特征*w*（例如：地理特征——四川盆地）和主题类*C*（例如：重要气候信息）之间的独立性。表示除*w*以外的其它特征，表示除*C*以外的其它主题类，那么特征*w*和主题类*C*的关系有以下四种情况：(*w*,*C*), (*w*,), (,*C*), (,)，用*A*, *B*, *C*, *D*表示这四种情况的文档频次，总的文档数*N*=*A*+*B*+*C*+*D*，统计量的计算方法如公式（2-1）所示。

 (2-4)

当特征*w*和主题类*C*之间完全独立的时候，统计量为0。统计量是归一化的统计量，但是它对低频特征的区分效果也不好。

### 2.1.3 特征提取实验步骤

本文在选取词特征和短语特征的基础上，对气象服务文本进行二元邻接词组分析。它是用于分析文本库中的相邻两个词的关系。其中某两个词在文本库中所一起出现的次数，称为共现频次。在前一个词出现的情况下，某词会出现在它后面相邻的概率，称为二元词组的转移频率。这种文本分析对业务员的写作习惯是非常有效的。

研究使用专门的文本分析处理工具（如图2-1所示），对气象语料库中所有的文本进行文本分词处理，然后对所有的分词结果进行词频统计和二元词组邻接分析。

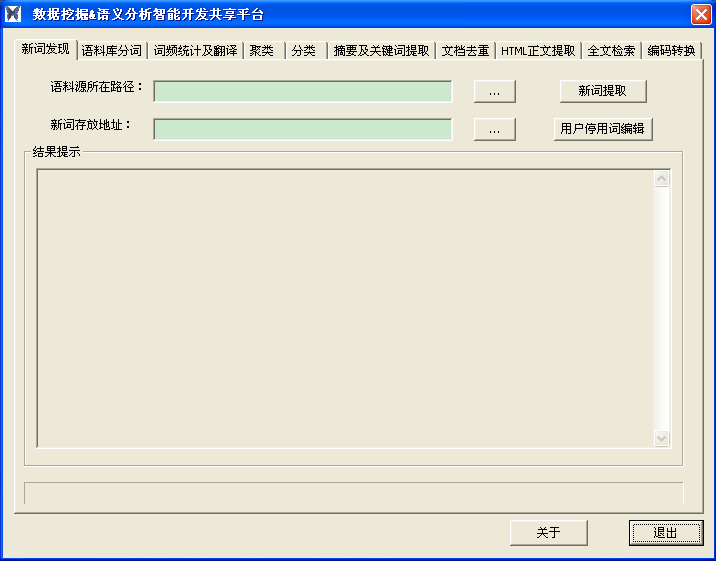


图2-1 文本处理界面

统计实验中，主题*C*是要提取特征的语料库文本，是被对比的语料库文本，*w*是要分析的特征词语。首先选择要对比和分析的语料库，然后将语料库的全部文本进行分词处理，对分词结果进行词频统计和二元邻接词组统计，并按照公式来计算特征值，对特征值计算结果进行排序，最终选出文本的特征词语。

算法：文本特征词提取算法。使用统计量方法，提取气象文本特征词。

输入：

C：主题类文本（天气预报、重要气候信息和每日天气提示）；

*w*：主题类文本中的特征词；

**：表示除*C*以外的其它主题类文本（人民日报）。

输出：主题类词语特征值计算结果。

方法：

Input：C和**的各自的词语总数和全部词语信息。

for C中每个*w*

｛

A = (*w*, *C*)；// 主题文本中*w*的个数

C = (,*C*)；// 主题文本中不含*w*的词语总数

for **中每个*w*

｛

B = (*w*,)；// 非主体文本中*w*的个数

D = (,)；// 非主题文本中不含*w*的词语总数

｝

N = *A*+*B*+*C*+*D*；

 // 统计量的值

｝

qsort( key: 主题C中所有*w*的统计量的值)；// 降序

Output：主题C中*w*的统计量值的降序表。

## 2**.2 特征与用语分析**

研究首先将天气预报信息、重要气候信息、每日天气提示跟人民日报文本作对比，在于得到气象服务文本各部分的用语特征，并从特征词中分析气候用语的变化。然后把重要气候信息以不同的要素划分，分析“降雨”、“风”等文本特征。如图2-2所示。



图2-2 特征提取结构图

### 2.2.1 地理区划特征

将天气公报信息、重要气候信息和每日天气提示与人民日报作对比分析，提取了这三类气象服务文本的地理区划用语特征词，如图2-3所示。



图2-3 地理区划特征词

如2-3左图所示，“内蒙古、东北、江南、华南、华北”和“西藏”的卡方统计量最高。说明预报员在编写天气公报时，一般按照一级区划进行天气描述。而中图和右图中只有很少的一级区划词语，更多的是行政区域和省份，如 “广西”、“贵州”等。由上图可以得出，天气公报文本中，“内蒙古地区”、“东北地区”等一级行政区划较多，而重要气候信息文本和每日天气提示中，行政区划的地理词语出现的较多，更能代表这两种服务文本的地理用语特征。在天气公报文本中，“四川盆地”的卡方统计量也很高，而行政区划“四川”出现的并不多，说明预报员描述天气的时候往往使用“四川盆地”。事实上，四川盆地是指四川东部的盆地，因其出现暴雨和地质灾害较多，故预报员对“四川盆地”的关注度远高于“四川”。从服务文本的二元邻接词分析中可以得到，文本历史库中“新疆山口”这个词一共出现了580次。其中，当“新疆”这个词出现的情况下，后面出现“山口”这个词的概率大约有5.9%，远大于一般地理要素词组的出现概率，而事实上，一旦有冷空气过境，“新疆山口”由于地形等原因经常容易出现大风现象，所以被业务人员经常使用。

一般来讲，预报员在描述气象要素空间分布时，尽量采用全国气象地理区划[27]（四级区域），但是通过文本的上述分析可以看出，预报员多年来形成的语言表达习惯，并不完全遵循全国气象地理区划。因此，本文对地理要素的统计分析与识别，将对空间区划的合理划分以及高质量文本公报的自动生成产生良好的指导意义。

### 2.2.2 地理方位和范围特征

气象服务文本的方位词如图2-4所示。三类气象服务文本中，“东部”出现的次数最多，“西部”出现的次数最少，“东南部”、“东北部”等词也经常被预报员所使用。基于自然语言统计和特征提取的结果可以推断出，预报员在描述某一个地方的时候，可能将地区按象限来划分，即“东”、“南”、“西”、“北”、“东南”、“东北”、“西南”和“西北”八个象限。



图2-4 方位特征词

如果预报员知道某种天气过程肯定会发生，但不能很准确地把握到底是地理区划的哪个部分，就会用到模糊用语[28]。本文提取到的表示地区大小的范围词主要有“大部”、“部分”和“局部”三个词，如图2-5所示。当被描述地区发生特定天气现象的象限个数达到某一阈值时，业务人员可能会根据自身经验进行判断，使用这三个范围词语。因此，范围词所映射空间大小的准确与否，将直接决定了该词是否被恰当的使用。选择范围词的方法有多种，可以采用面积法（根据空间所占的比例来确定范围词）等。还有学者提出了采用Sugeno积分语言量词模型等[29]，这些关键技术的解决将对气象服务文本的自动生成奠定良好的基础。



图2-5 范围特征词

### 2.2.3 气象要素特征

**总体特征:**“天气公报信息”、“重要气候信息”和“每日天气提示”这三类气象服务文本关注天气现象的侧重点不同。将这三类与“人民日报”作对比分析，提取天气现象用语特征词，如图2-6所示。



图2-6 天气现象特征词

“天气公报”中，气象词语种类繁多，而且词频较高，如图3-4的左图所示。在出现最多的前20个气象词中，包含“雨”的词语有10个，表明天气公报中描述最多的天气类型是雨。上图中部是“重要气候信息”文本的气象特征词语。特征词中出现“暴雨”和“降雨”，更多出现的则是“台风”、“热带风暴”、“沙尘”等灾害性天气，这说明它描述的多是重大气候过程；“每日天气提示”气象名词特征值表可以看出，“降雨”排在第一位。“冷空气”、“降温”、“暴雨”和“雪”等名词表述了每日天气提示最经常出现的天气状况。“台风”、“热带风暴”、“阵风”、“大风”和“风力”也在一定程度上代表了每日天气提示文本用语的特征。

**“降水”描述特征:**研究提取了“重要气候信息”文档中描述“降水”的服务文本，进行词频及二元邻接词分析可知：“干旱”、“毫米”和“降水量”这三个描述降水特征的词出现的频率较多。由自然语言的统计结果可获得常识性知识，即预报员描述“降水量”的单位通常是“毫米”；当某一地区的降水量达到最大值或最小值的时候，预报员习惯将这个极值描述出来，而“降水量”位于中间区域时，则按照降雨的量级（“小雨”、“中雨”、“大雨”“暴雨”等）来预报；此外，“气象”和“干旱”这个二元词组出现的次数也很高，而且转移概率接近为1，同时，“降水”服务文本中，气象要素“干旱”出现的次数最多，并且卡方统计量的值也最大（如图3-5所示）。事实上，“干旱”本身属于气候概念，同时干旱监测也是气候中心的业务，而“高温”、“暴雨”更多是天气的范畴。故近年来“重要气候信息”文档中以“气象干旱”为主题的文本占了很大的比例。

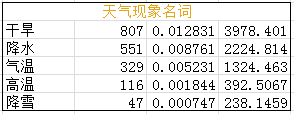


图2-7 “降水”主题文本特征词

**“台风”描述特征:**“台风”主题包括“重要气候信息”文档中描述“台风”天气状况的文本。通过词频分析和二元邻接分析可知，气象词“台风”和“热带风暴”以及地理词“福建”和“广东” 词频最高。“经济”和“损失”这个二元词组出现的次数也很多，而且转移概率为1。同时，“受灾”、“灾害”、“严重”、“死亡”和“失踪”词频也很高，说明此类天气现象往往造成经济损失与人员伤亡。

图2-8为“风”主题文本的部分特征词提取结果。“台风”是最能代表此类主题天气现象的特征词。地理名词“福建”、“广东”、“浙江”和“海南”表明了受灾较多的地区为我国的东南沿海地区。“菲律宾”、“越南”则反映了某些风暴在登陆我国之前的路线图。故而可以推断出热带气旋（“台风”等）往往在中国东南沿海登陆，严重影响当地人民的生产、生活。



图2-8 “风”主题文本的卡方统计量分析

**其它特征:**研究将“重要气象信息”文档按照签发人分类，并进行分析。图2-9左侧为独立签发人的比例图，饼状图可以清楚地反映前6位所占比例超过50%。图右侧是前6位的独立签发人在2008-2011年间所核定和签发的数据比例。从数据的分布中可以看出, 每个独立签发人的签发次数逐年递减，并且都在2011年到达谷底，也表明了气候部门的业务分工也越来越密切，逐步形成了主班、领班、及首席的共同签发的模式。

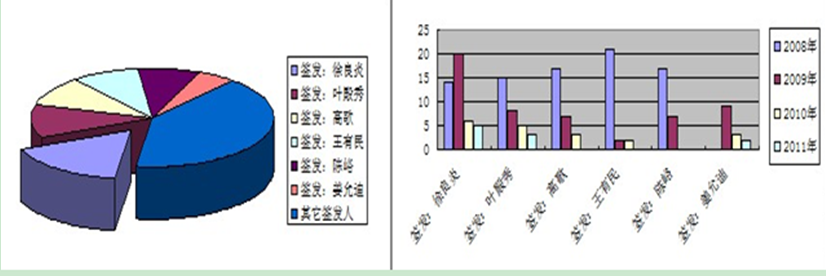


图2-9 签发人数据分析

图2-10为“徐良炎”签发文本的特征词分类结果。由图右侧可知，“干旱”出现的次数最多，说明了“徐良炎”签发这类服务文本居多，这也意味着“徐良炎”是干旱领域的气象专家。

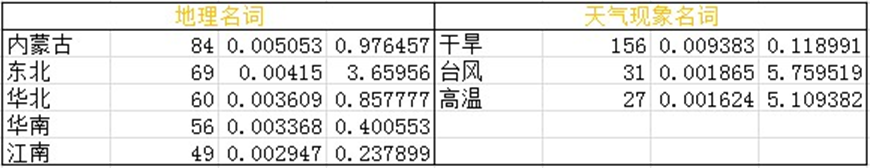


图2-10 徐良炎签发文本特征分析

## 2**.3 三类服务材料用语分析**

通过本文分析，可以看出不同类别的服务材料之间差距很大。“天气公报”出现的地理区划名词较多，气象要素描述词也非常全面，它着重在于整体的描述全国的天气情况，文本结构单一。文本结构包括时间描述、地理区划描述和天气现象描述三部分。地理区划描述中，既有象限划分的词（如“河南南部”），也有模糊量词（如“湖北大部”）。而天气现象描述中，天气现象的量级区分出现较多。

而“重要气候信息”、“每日天气提示”文本的结构复杂，且服务主题多变，没有固定的模版。重要气候信息描述的多为灾害天气及过程，如“洪涝”、“干旱”、“沙尘暴”、“高温”以及“台风”等，提供决策服务；而每日天气提示则是将专业的气象语言转换成通俗易懂的语言形式，以服务群众生活出行，要求用词简单、易懂、甚至可以用词鲜明、修辞和口语化。

本章研究的不同气象服务类型习惯用语和句子行文结构为气象文本的计算机自动生成技术以及气象文本的自然语言表达奠定良好的技术基础。

# 第三章 GIS空间分析与气象原始数据生成

上一章介绍了自然语言统计学的研究方法，以气象部门（气象、天气等领域）人工撰写的典型文本作为对象进行了特征抽取和综合分析，为气象自然语言模型建立了基础。实际系统运行的时候，输入的数据类型为Shapefile，不能直接进行数据变换、分析，更不能作为自然语言处理的初始数据。本章针对实际出现的问题，把多维、复杂的地理原始数据与气象数据进行匹配，并详细的定义了后续实验所需要的标准气象数据格式。

## 3.1 GIS空间分析概述

随着科技的发展，气象观测数据也包含了越来越多的信息，这种信息包括从分布各地的观测站得到的温度、风向、降水量等气象数据，这些数据对于抗旱防涝、生产生活、环境治理、水土保护诸多领域具有极其重要的参考价值。气象数据中的地理、风向等都与地理信息密不可分，因而可以说，气象数据严格意义上也属于地理空间数据。气象数据的处理分析与地理信息系统的空间分析技术密不可分。

### 3.1.1 地理信息系统

在现阶段的人工智能研究领域中，自然语言理解、自治机器人导航、物理位置常识推理、图像理解、计算机视觉等方向涉及地理信息系统的空间分析技术。

地理信息系统（Geographic Information System）[30] 是为地理研究和地理决策服务的计算机技术系统，它是计算机科学、地图学、测量学和地理学等多门学科的交叉，在依靠地理空间数据库的基础上，采用地理模型的分析方法，给人们提供多种动态的、空间的地理信息。

我国地理信息系统起步比西方国家来说相对较晚，但经过我国研究人员10年的努力，这方面的理论和技术体系也越来越成熟和完善，并有着向多功能综合性系统发展的趋势。在经济、社会、生活等诸多方面，GIS技术应用的广度越来越宽泛、深度也不断的加强。由最开始的简单绘制静态电子地图到日渐完善的动态监测和分析[31]；由单一的地理数据管理到规范的规划辅助决策等。GIS查询和管理气象信息非常方便，在海量信息处理方面更是有优势，在分析处理地理空间数据的同时，又可以与数值、图形的表示相结合，能够形象直观地表示出数据结果。GIS的空间分析能力非常强，能够提供动态的、空间的地理信息，为气象信息决策服务提供科学依据。

GIS空间分析的主要方向是综合处理地理空间信息，在解决某些特别复杂问题时具有强大信息处理能力。空间分析的技术方法主要有：（1）基于地图的空间分析技术，如叠加分析、缓冲区、地学图谱等；（2）地理空间动力学分析，如空间价格竞争模型、择位模型等；（3）地理信息空间分析，融合不同专业领域的空间分析模型。

### 3.1.2 地理空间数据库

在经济、文化、贸易、知识全球化的时代，地理空间信息资源逐渐成为了当今社会战略性信息基础资源不可缺少的一部分，地理空间信息产业也为知识经济的发展提供了源源不断的动力。其中，地理空间信息技术的自主创新主要体现在开源地理信息空间分析技术的发展加速。实验需要把复杂的地理信息转换为标准格式的气象初始数据，这就需要GIS开源项目来进行空间分析。

本文选定了开源GDAL(Geospatial Data Abstraction Library)、GEOS作为地理信息空间分析的工具。GDAL/OGR是一个在X/MIT许可协议下的开源栅格空间数据转换库，它使用抽象数据模型来表达项目所需要的全部文件格式，OGR是GDAL工程的一个部分，OGR提供对矢量数据的读写支持。GEOS的全称是“Geometry Engine, Open Source”，它提供了Open GIS Consortium规范中简单几何要素对象操作的C++语言的实现。GEOS包含了一个集合形状的拓扑关系操作实用库，也可以说它是判断不同几何形状关系和对不同几何形状操作以及生成新几何形状的操作库。拓扑模型是地理信息系统领域最重要的一个组成部分，其计算方法虽然简单但是实现起来却很困难，OGR提供了拓扑模型计算的规范接口，而这些规范的接口在GEOS中被完美实现。与其他GIS项目相比较，GEOS在“空间谓词”和“空间操作”这两个方面有非常大的优势。GEOS支持的空间关系计算如表3-1所示。

表3-1 GEOS空间关系计算

|  |  |
| --- | --- |
| 相等(Equals) | 几何形状拓扑上相等 |
| 脱节(Disjoint) | 几何形状没有共有的点 |
| 相交(Intersects) | 几何形状至少有一个共有点 |
| 接触(Touches) | 几何形状有至少一个公共的边界点，但是没有内部点 |
| 交叉(Crosses) | 几何形状共享一些但不是所有的内部点 |
| 内含(Within) | 几何形状A的线都在几何形状B内部 |
| 包含(Contains) | 几何形状B的线都在几何形状A内部（区别于内含） |
| 重叠(Overlaps) | 几何形状共享一部分但不是所有的公共点，而且相交处有他们自己相同的区域。 |

GEOS空间叠加分析操作能力也很强，主要有以下几个操作，如表3-2所示。

表3-2 GEOS空间叠加分析操作

|  |  |
| --- | --- |
| 缓冲区分析(Buffer) | 包含所有的点在一个指定距离内的多边形和多多边形 |
| 凸壳分析(Convex Hull) | 包含几何形体的全部点的最小凸壳多边形 |
| 交叉分析(Intersection) | 多边形AB中所有共同点的集合 |
| 联合分析(Union) | AB的联合操作即为AB全部点的集合 |
| 差异分析(Difference) | AB形状的差异分析即为A里有B里没有全部点的集合 |
| 对称差异分析(Symmetric Difference) | AB形状的对称差异分析即为位于A中或者B中但不同时在AB中所有点的集合 |

## 3.2 气象数据接口规范

在进行GIS数据转换、分析之前，要定义气象原始数据的接口规范，以避免因数据格式不规范而带来的结果错误，好的接口规范可以提高实验结果的精确性。定义的XML接口里面规范了系统中各种特定信息的数据格式。主要结构如图3-1所示。



图3-1 XML接口数据格式定义

XML结构的根节点为GIS\_INFO，其中包含了一天或者几天的气象信息。每一天的气象信息只包含Basic\_Info和Include\_Info。其中Basic\_Info是某一地理区划的气象信息，Include\_Info则是各种气象落区的相互包含关系。

如图3.1所示，在Basic\_Info中，有两种格式的Item，这两种Item共同拥有的属性值为：ID即节点的唯一标号，Geo\_name即地理要素的名称，Type即气象的大类型，Weather\_Name即为Type中细化的天气现象，Value为气象类型对应的标号值。气象领域中，主要关注的是“风”和“降水”这两种气象特征，第一个Item类型为“风”，与“降水”不同的是，这类的Item里面有“Direction”属性，即“风向”；在“降水”类型的Item中，与之不同的属性有：“Area”即特定气象覆盖的面积，“Proportion”即地理要素发生特定天气现象占地理要素总面积的百分比。

“降水”类型的Item如图3.2所示，样例中，展现的是标志Id为9的降水圈，细化到具体的Weather\_Name就是“小雪”。在气象中，“雨”跟“雪”都是属于降水天气，在气温低于零度后，降水就变成降雪。在Id为10的降水圈中，Type是“雨雪”，而与Id为9的东北地区不同，西藏地区的Weather\_Name就是“小雨”。东北地区的雨雪圈的面积是526839平方公里，小雪天气发生的占整个东北地区的66.8%。

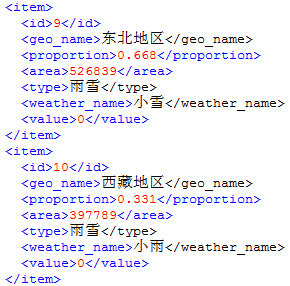


图3-2 “降水”类型的Item

“风”类型的Item如图3-3所示，描述气象的落区为内蒙古西部，当Type为“风”的时候，Weather\_Name的含义是风的级数，从图中可以看见，内蒙古西部的风级数为6-7级。

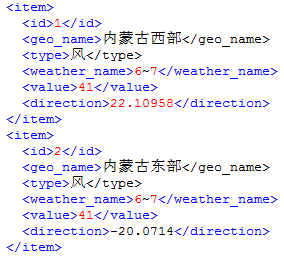


图3-3 “风”类型的Item

图3-3中的Direction为实际风向的角度，在规范中，采取8个象限来描述风向，如图3-4所示，Direction的角度为与X轴的夹角，若Direction的角度为正，则是将X轴逆时针旋转相应的度数，反之，则将X轴顺时针旋转相应大小的度数，通过最后的风向的落区来判断风的方向。



图3-4 风向角度的判定

Include\_Info的节点中表示的是降水圈包含关系，Id\_Big表示某个特定气象的一个大圈，Id\_Small则表示在这个大圈包含内的某个小圈。如图3-5所示，Id为5的内蒙古地区的降水圈包含了id为25的内蒙古西部的降水圈。

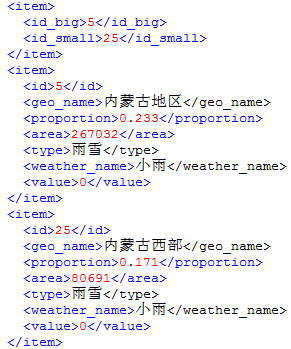


图3-5 地理要素包含关系

## 3.3 气象服务文本的空间推理分析

目前，许多学者提出了不同的空间分析推理方法，但这些地理空间推理的方法大多基于代数以及逻辑[32-33]。如何为地理空间信息推理添加语义支持，如何确定地理空间信息的规范表达和推理规则的正确推导方法，是研究迫切需要解决的问题。

气象观测站观测的实时数据与地理空间信息密不可分，这种数据包括与地理空间分布有关的所有要素，具有很大的灵活性、伸缩性[34]。地理空间信息根据规划属性可以简单的划分为行政区、商业区、居住地、耕地、工业的平面结构，也可以依照生态环境广义地划分成人群、资源、生态等信息，更能详细的、分类别的讨论河流、公路、山脉等的具体情况。因而在进行空间分析推理时，需要确定范围，明确目标，以及领域涵盖的内容，甚至要涉及层次内容等级、领域范围等。

本文需要从多方面指标（方向、位置、距离、尺度）等多方面指标来提取气象空间信息的特征。GIS的空间分析方法有：空间缓冲、空间插值、分类分析、聚类分析等，运用这些方法可以提取暴雨级、特大暴雨级、沙尘暴类似的描述空间的特征，例如可以直接把观测站的数据进行插值形成气象信息的空间分布场，或者取出单站峰值来表达不同气象要素（温度、风场、湿度等）。而在提取更有效、准确的气象空间位置特征时，需要建立更完善的地理区划方法。结合不同类型的气象服务文本特征，在建立区划时，要考虑以下问题：（1）如何保证落区的范围描述的合理、精确性，即使用类似东北地区级还是黑龙江省级的问题；（2）行政区域、地质区域、经济区域、交通区域的结合问题。本文将进一步分析，在实际的实验中，解决上述问题。

### 3.3.1 气象地理要素

本文对2008-2011年由国家气象局发布的气象服务文本进行地理要素的统计分析，选取了这三年来业务人员使用的全部地理要素名词。结合气象公报的语言描述特点，将提取的地理要素的名词分成四级：

1. 一级地理区域：西藏地区、内蒙古地区、新疆维吾尔自治区、东北地区、西北地区、华北地区、华南地区、黄淮地区、江淮地区、江汉地区、江南地区和西南地区；
2. 二级地理区域：在一级要素基础上划分，例如西藏南部、黄淮地区北部、内蒙古东部等；
3. 三级地理区域：北京市、上海市、天津市、重庆市、河北省、河南省、山东省、山西省、辽宁省、吉林省、黑龙江省、浙江省、江苏省、安徽省、江西省、陕西省、贵州省、云南省、湖北省、湖南省、福建省、广东省、海南省、四川省、甘肃省、青海省、台湾省；
4. 四级地理区域：在三级地理要素区域基础上划分，例如河南北部、青海南部等。

综上所述，一级气象地理要素的变量词有12个，其中西藏地区、内蒙古地区、新疆维吾尔自治区既是一级气象要素，也是行政规划。全部的一级地理要素区域组成了中国疆域。二级地理区域在一级区划的基础上划分的。中国各个行政省份、直辖市构成了三级区划，四级则是在三级的基础上更细划分得出的、三四级划分的关系与一二级相似。

### 3.3.2 气象地理区划

虽然气象服务文本中的区划大致分为四级，但为了更好的适应气象服务精细化的要求和具体的观测站点分布限制，每一个区划被细分的方法也不尽相同。例如东北地区划分为东北东部、东北西部，而西藏地区却划分为西藏东部、西藏南部、西藏北部、西藏西部。

为此，本文构建了上述四级地理区划的具体包含关系，标准如下：

1. 二级 ⊆ 一级；
2. 三级 ⊆ 一级；
3. 四级 ⊆ 三级；
4. 二级、三级之间无包含关系。

由标准可以得出，一级区划包含了二级、三级区划，二三级区划并无直接的联系，一级与二、三两级的具体包含关系如表3-3所示。

表3-3一级与二、三级地理区划关系表

|  |  |  |
| --- | --- | --- |
| 一级地理区划 | 二级地理区划 | 三级地理区划 |
| 东北地区 | 东北东部、东北西部 | 吉林省、辽宁省、黑龙江省 |
| 西北地区 | 西北地区东部、西北地区中部 | 青海省、陕西省、甘肃省、宁夏回族自治区 |
| 江汉地区 | 江汉东部、江汉西部 | 湖北省、河南省 |
| 华南地区 | 华南东部、华南西部 | 广西壮族自治区、广东省、福建省、台湾省、海南省 |
| 黄淮地区 | 黄淮东部、黄淮西部、山东半岛 | 河南省、江苏省、安徽省、山东省 |
| 江淮地区 | 江淮东部、江淮西部 | 河南省、安徽省、江苏省、湖北省 |
| 西南地区 | 西南地区北部、西南地区东部、西南地区南部 | 云南省、贵州省、四川省、重庆市 |
| 华北地区 | 华北东部、华北西部 | 山西省、河南省、河北省、山东省、山东半岛、北京市、天津市 |
| 江南地区 | 江南东部、江南西部 | 湖北省、江西省、浙江省、福建省、安徽省、江苏省、上海市、湖南省 |
| 内蒙古地区 | 内蒙古东部、内蒙古中部、内蒙古西部 | 内蒙古地区 |
| 西藏地区 | 西藏北部、西藏东部、西藏南部、西藏西部 | 西藏地区 |
| 新疆维吾尔自治区 | 新疆东部、新疆西部、新疆南部、新疆北部 | 新疆维吾尔自治区 |

从标准中可以得出四级区划属于三级区划，其具体的包含关系如表3-4所示。

表3-4 三级与四级区划包含关系表

|  |  |
| --- | --- |
| 三级地理区划 | 四级地理区划 |
| 北京市 | 北京市 |
| 天津市 | 天津市 |
| 上海市 | 上海市 |
| 重庆市 | 重庆市东北部、重庆市东南部、重庆市西部、重庆市中部 |
| 台湾省 | 台湾北部、台湾东部、台湾南部、台湾中部 |
| 新疆维吾尔自治区 | 新疆北部、新疆东部、新疆南部、新疆西部 |
| 吉林省 | 吉林省东部、吉林省西部、吉林省中部 |
| 河北省 | 河北省北部、河北省南部、河北省中部 |
| 黑龙江省 | 黑龙江省北部、黑龙江省东部、黑龙江省南部、黑龙江省西部 |
| 甘肃省 | 甘肃省东部、甘肃省河西地区、甘肃省南部、甘肃省中部 |
| 广东省 | 广东省东北部、广东省东南部、广东省西北部、广东省西南部、广东省中部 |
| 海南省 | 海南省东部、海南省南部、海南省西部 |
| 山西省 | 山西省北部、山西省南部、山西省中部 |
| 辽宁省 | 辽宁省北部、辽宁省东部、辽宁省南部、辽宁省西部、辽宁省中部 |
| 陕西省 | 陕西省北部、陕西省南部、陕西省中部 |
| 宁夏回族自治区 | 宁夏北部、宁夏南部、宁夏中部 |
| 青海省 | 青海省北部、青海省东部、青海省南部、青海省西部 |
| 山东省 | 山东省北部、山东省东部、山东省南部、山东省中部 |
| 安徽省 | 安徽省北部、安徽省南部、安徽省中部 |
| 河南省 | 河南省北部、河南省东部、河南省南部、河南省西部、河南省中部 |
| 江苏省 | 江苏省北部、江苏省南部、江苏省中部 |
| 四川省 | 四川省北部、四川省东部、四川省南部 |
| 江西省 | 江西省北部、江西省南部、江西省中部 |
| 浙江省 | 浙江省北部、浙江省南部、浙江省中部 |
| 湖北省 | 湖北省北部、湖北省东部、湖北省南部、湖北省西部 |
| 福建省 | 福建省东北部、福建省东南部、福建省西北部 |
| 云南省 | 云南省东北部、云南省东南部、云南省西北部、云南省西南部、云南省中部 |
| 湖南省 | 湖南省北部、湖南省南部、湖南省西部、湖南省中部 |
| 贵州省 | 贵州省东北部、贵州省东南部、贵州省西北部、贵州省西南部、贵州省中部 |
| 广西壮族自治区 | 广西壮族自治区东北部、广西壮族自治区东南部、广西壮族自治区西北部、广西壮族自治区西南部 |

## 3.4 气象原始数据生成

本节解读地理输入的初始数据的每个组成部分，介绍了读取、转换以及生成规范的XML接口数据的方法。

### 3.4.1 地理原始数据

本文采用的地理数据格式为Shapefile，Shapefile文件是GIS系统格式文件，由美国环境系统研究所设计，其文件的数据集中存有空间特征表中的属性信息和非拓扑集合对象。Shapefile特征表中的集合对象由坐标点集来表示的图形文件构成。Shapefile是符合工业标准的矢量文件，不含拓扑数据结构，主要由三个文件构成：

主文件(.shp)，可直接存取，其中每个记录都描述了一个地理特征的全部坐标值，由固定长度的文件头和不定长度的空间数据组成。固定长度的头文件占100个字节，内容包括了含空间数据的元数据，如文件的长度、图层的范围、Shape的类型等。读入文件头获取Shape文件的基本信息是导入空间数据的先决条件，并在此基础上建立对应的元数据表。不定长度的空间数据是由定长的记录头和变长度记录内容组成，且每条记录都由记录头和空间坐标对组成，其记录头由记录号（从1开始）、坐标记录长度（16位字来衡量）这两个记录项组成。

索引文件(\*.shx)，主要包含坐标文件的索引信息。文件中每个记录包含对应的坐标文件记录距离坐标文件的文件头的偏移量，故可以快速的定位到指定目标地的坐标信息。索引文件由文件头和实体信息两部分组成，其中文件头部分包含100 字节的记录段。

属性表(\*.dbf)，包含相对应主文件中的全部地理特征的属性，属性表中的记录与主文件中的记录顺序完全相同，根据表中的索引号来确定几何记录和属性数据之间的一一对应关系。属性表是一个标准的DBF文件，由头文件和实体信息两部分构成。其中文件头长度不定，是关于DBF的总体说明。头文件主要内容是对所有记录项信息的详细描述，例如对名称、长度、数据类型等。需要依次循环来读取每条记录。

在矢量矩阵数据中，所有的雨、雪、霜等气象现象都是由各种不规则的闭合多边形叠加而形成的，图3-6是国家气象局某日从全国各个观测站收集到的气象数据的部分截图。图中可以看到一个N×3的规模的数据矩阵，矩阵开头的内容是气象时间和降水量等信息，每一行的三个数据（经度、纬度和海拔）都组成了一个点的全部信息，而实际上气象公报并不需要考虑海拔因素，故海拔的数据默认是0，全部点的集合就构成了一个闭合的不规则多边形。

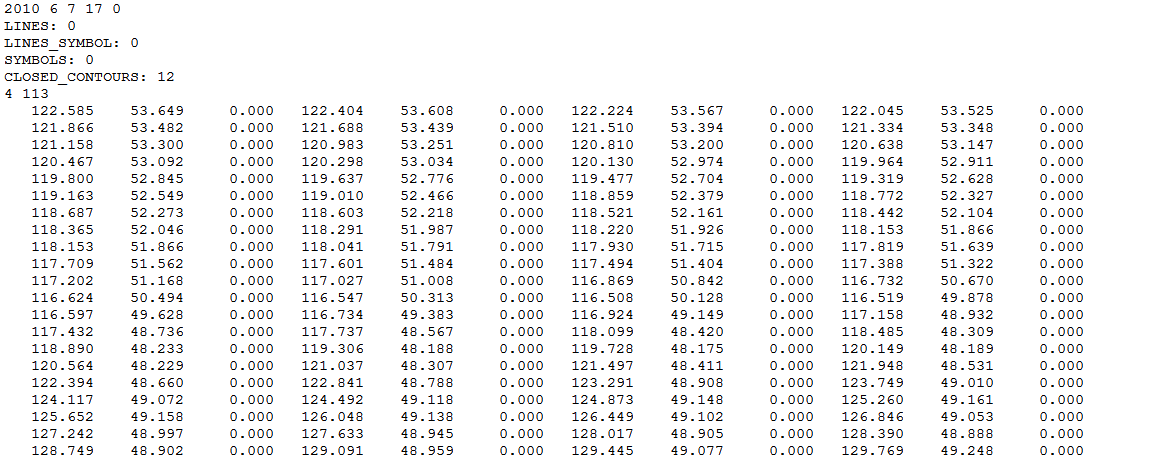


图3-6 地理原始数据矩阵

### 3.4.2 地理数据分析与处理

图3-6所示的是2011年04月11日的气象云图，图中每一降水量线的区域颜色都不相同。在实际处理过程中，每个降水量线都是一个闭合的不规则多边形，而正是全部的多边形经过叠加和着色形成了气象云图。

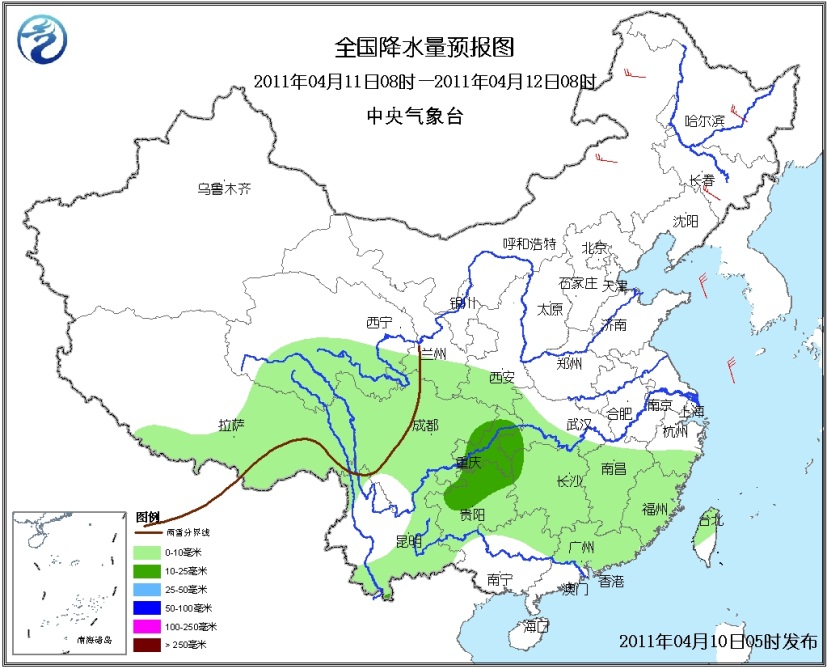


图3-6气象云图

系统实时的接收全国气象观测的数据汇总，在规定时间内，对降水圈的范围进行分析与提取。如图3-7为提取某日全国范围的降水圈图。每一个闭合多边形都是一个降水圈，如果两个多边形有叠加，很有可能就是两个叠加多边形的气象类型相同，而重叠的部分表现更为显著。例如，大圈的气象现象为小雨，重叠的小圈内可有可能是中雨，或者暴雨。

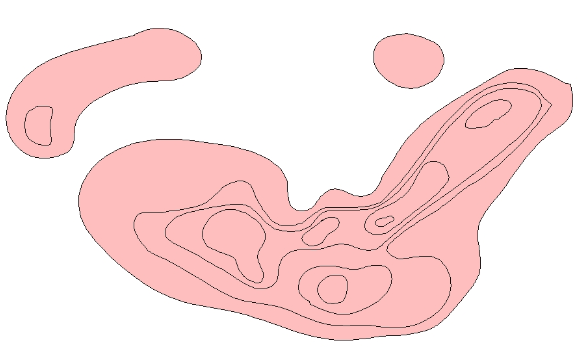


图3-7 降水雨圈

每一个降水圈都有相对应的属性值，图3-8为降水圈的属性表，FID就是标识号，形状都是闭合的多边形，标号值就是天气现象的类型。实验中主要提取这三类数据。原始数据中，标号值的每一个数值都代表了一种天气现象。

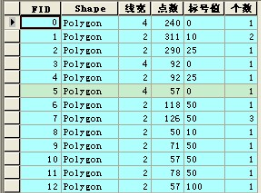


图3-8 降水属性表

本文将提取的标号值替换成相应的气象现象，如标号值为0的气象现象替换为“小雨”、标号值为100的气象现象替换为“大暴雨”等。雨雪标号值的全部含义如表3-5所示，由于气象公报中主要关注的是降水，风的标号值的对应表在本文中不再赘述。

表3-5 降水标号值

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 标号值 | 0 | 10 | 25 | 50 | 100 | 250 | 5 | 2.5 |
| 气象现象 | 小雨 | 中雨 | 大雨 | 暴雨 | 大暴雨 | 特大暴雨 | 大雪 | 中雪 |

在提取气象云图的所有信息之后，本文运用GIS空间分析方法来将气象要素与地理要素一一对应。由上节可知，系统采用了四级地理区划图，实验中，需要将每一级的地理区划矢量文件与气象落区做相交分析。



图3-9 地理三级区划图

图3-9为地理三级区划图，与降水圈的形状类似，每一个行政省份在矢量文件里都是一个闭合的多边形。具体实验中，将每一个多边形与降水圈的多边形进行相交分析，重叠的部分就是发生特定天气现象的地方，相交的面积就是XML接口规范中的Area的值，将这个值与地理要素的面积相除，就得到了Proportion的值。图3-10为相交的结果图。

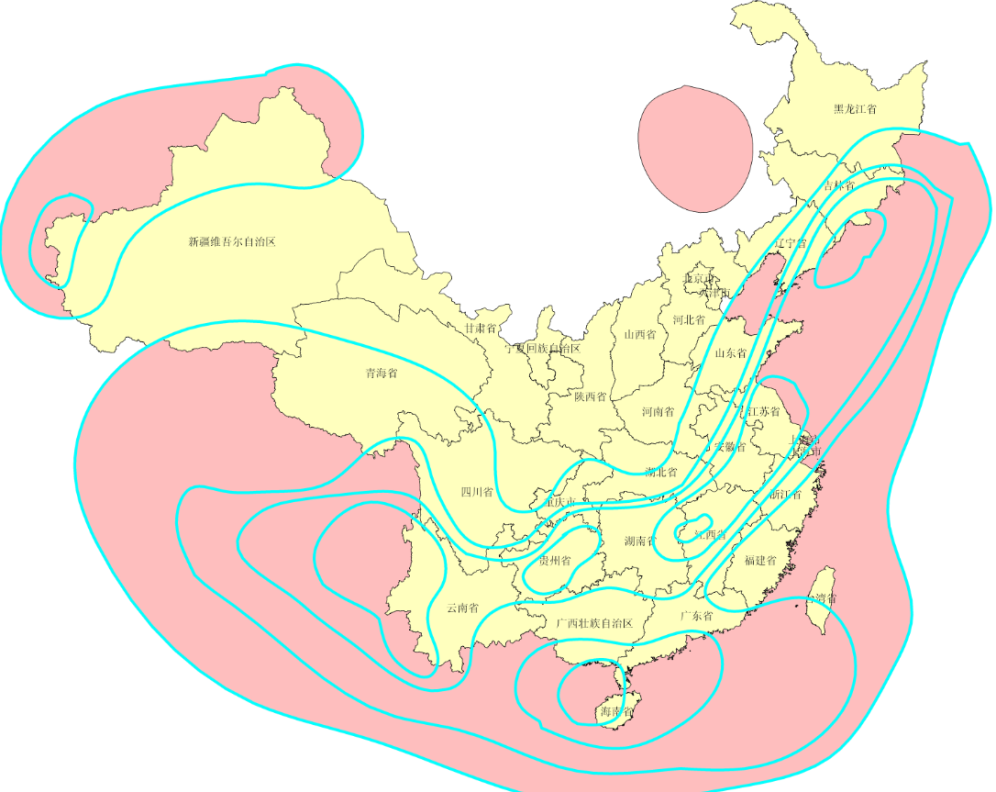


图3-10 空间分析结果

本文截取了部分空间分析的结果，如图3-11，图右FID为5的降水圈与北京市、山西省有交集，且气象标号值为0，这就说明北京、山西有小雨。同时，矢量数据需要通过空间投影转换的方法来与规范接口的数据规格保持一致。通过GIS空间分析方法，本文可以生成规范的、精确的气象数据，也为后续自动化生成的气象公报内容的合理性提供了数据方面保障。



图3-11 相交属性

# 第四章 气象公报文本生成

通过上一章介绍的GIS空间提取方法，实验已经得出标准规范的接口数据，在此基础上，本章对气象原始数据进行了更深一步的分析，建立了语言生成模型，介绍了基于NLP的自动化生成气象公报的详细步骤。

## 4.1 气象公报结构分析

本节主要研究的内容是句子的组织结构，段落的内容排列顺序，以及地理要素、气象要素的结构化描述方法。在气象领域的语言生成过程中，特别需要提取气象发生的地点、时间、类别等，并转换为框架模型来组成文本。

国家气象中心发布的气象公报主要由卫星云图、文本公报这两个部分组成。本文重点就是根据实时的卫星云图对应的数据，转换分析得到符合公众阅读习惯的、符合业务人员写作规范的文本。

气象公报由时间、降雨以及风向等描述内容组成。为了更清晰的分析气象公报的结构，对公报不同的部分设置不同的颜色。图4-1所示为任意选择一篇气象公报进行结构化分析的结果。在分析结果中，用红色字体来标明时间，即8日08时到9日08时这一天的时间跨度，在文字生成时，可将日期作为时间变量来处理。紧随着时间描述的是降雨相关的信息，它可以由结构“[地理要素]有[雨雪要素信息]”来表示，风向信息由“[地理要素]有[级数]风”来表示。[地理要素]是上一章中提出的地理区划内容，在上图中由蓝、紫、橙色字体标识。雨雪、风等天气是有固定量级的，在描述文本的时候，要根据实际的数值来替换成相应的量级。

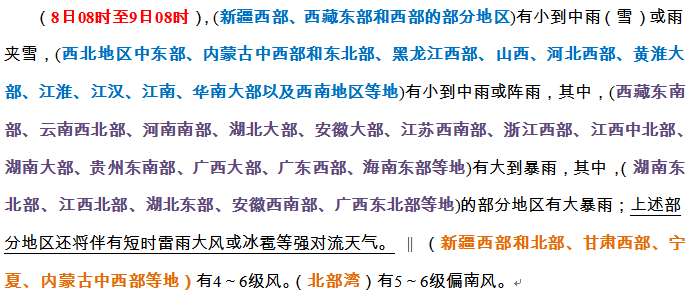


图4-1 天气公报结构图

## 4.2 跨领域的语言自动化生成模型

跨领域的语言自动化生成模型描述的是从GIS数据分析、转换到生成符合公众理解的语言文本的整个过程，是文本的合理组织和简洁表达[35-36]。本文首先要考虑到的是NLP现有的、成熟的技术框架[37-38]，也要引入复杂GIS数据的分析[39]、投影、文本生成等的建模方法。

针对气象公报的文本结构、用语特点以及需求现状，考虑到受限领域与通用领域相结合的现状，课题设计的跨领域的语言生成模型有如下特点：

1. 从框架匹配的模型来看，气象服务文本种类繁多，首先得从相应产品的历史数据库中学习、抽取结构和数据特征，建立相应的模版以适应文本生成。
2. 从预报规则来看，同一级地理区划的地方可能有相同的气象状况，但是在预报时，气象中心有着固定的预报优先级顺序。
3. 从句子优化来看，要除去繁琐、冗余的描述。例如在文本生成的时候，“……有小雨，….有中雨，……有4级东风，……有6级东风”往往可以优化为“……有小到中雨，……有4-6级东风”。
4. 从句子规划来看，句子模版主要有雨雪、风和气温这3种类型。

本文气象服务文本的特征用语，在使用管道式体系结构[40]的基础上，进行了一部分修改和扩充，如对文本框架模型匹配、句子优化、预报规则、段落规划等多方面的处理。具体模型如图4-2所示。



图4-2 语言生成模型

## 4.3 受限领域的文本自动化生成

本文的气象语言生成模型主要由以下三个部分组成：

1. 短语规划模块，主要包括短语结构统计整理和气象、地理要素统计与规划两个部分；
2. 句子规划模块，主要包括读取规则库以及将气象数据的内容填入句子模版；
3. 篇章规划模块，在基于自然语言处理的原则上，调整篇章结构，去除冗余文字，最终形成准确的、易懂的气象公报文本。

### 4.3.1 短语规划

本文对历史气象服务文本进行了信息提取，主要抽取了地理要素短语、气象要素短语以及描述气象的短语结构模版。气象公报中完整的描述气象的句子也正是由要素和短语结构模版组成。

经过统计分析，近三年气象服务文本中用到的词语大约有2600个，并且有相当一部分只出现了几次。而研究的重点在于文本的通用历史规律，所以在统计词频后，去除出现次数较低的词语，并选取经常出现的高频词。

气象要素变量包括各种描述气象信息的专业用语，例如“东风”、“霜冻”、“暴雨”以及“台风”等。地理要素变量则包含上文提到的地理区划的内容。气象、地理要素变量具有两个共同的特征：数量有限且指代明确。

短语结构模板中包含了要素变量以及固定的搭配语句。例如模板句“[地理要素]有[气象要素]”。

### 4.3.2 句子规划

本文统计总结了气象文本中的短语模版以及各类变量的信息，并分别在风、雨雪、气温以及能见度这四个气象领域中建立了句子模板库。图4-3是模板库中的部分例子，每一个[]之中包含一个信息变量。下图中所示的信息变量有[方向]、[级数]、[地点]、[雨雪强度]等。

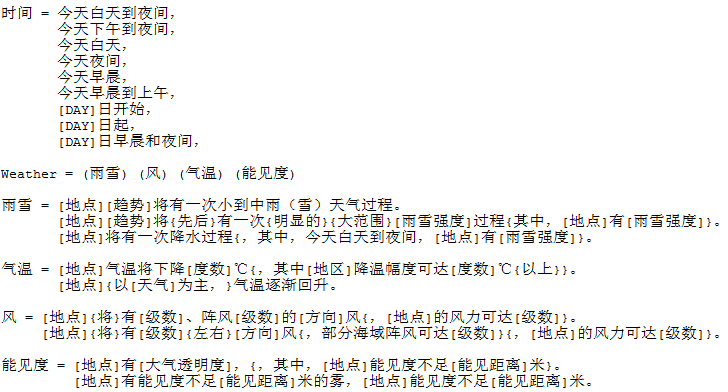


图4-3 气象文本模板库

在地理GIS数据经过空间分析之后，系统从标准规范的数据接口中读取气象信息，并根据Item中的Weather\_Name来确定气象类型并选择句子进行描述。系统实施了一个简单的“填槽”的处理，具体过程如下：系统取出适合待插入气象类型的句子模版，并进行空间推理分析，选择准确的、简练的描述词语，处理后替换到句子模板的变量中，并最终形成完整的描述气象的句子。

首先系统对Item中所有的地理信息遍历，判断每一个Item中的Proportion值，设为P，若P<=0.05，那么将此Item直接删除。系统将再次遍历所有的地理信息，对同一类型的天气现象所对应所有的Item中的地理信息提取出来，并对二级区划以及四级区划的地理信息进行合并处理。系统将属于同一个一级区划的所有二级区划Item中的面积累加，将累加值与对应一级区划的面积相除，若该值达到一定的阈值，则合并，反之则不做合并处理，对四级区划的区域处理采取与二级区划同样的方法，相应的阈值可以在系统中设定。

例如，在当Item里气象的类型是中雨时，选择的句子模版为“今天白天到夜间，[地点]将有[雨雪强度]。”。系统分析提取Item的地理信息，整合后可以得到简练的、准确的地理位置，即“内蒙古东部、内蒙古中部、内蒙古西部、西南地区北部、西南地区东部、黄淮东部等地”。如果不经过处理，直接填槽即代入变量的话，即的到“今天白天到夜间，内蒙古东部、内蒙古中部、内蒙古西部、西南地区北部、西南地区东部、黄淮东部等地将有中雨”。可以看出，地理要素信息过于冗余，需要进一步简化处理。通过查询地理区划可以得出，内蒙古地区由内蒙古东部、内蒙古中部、内蒙古西部这三个部分组成，可以合成为内蒙古地区，西南地区北部、西南地区东部可以合成西南大部。简化后得到的气象公报内容为“今天白天到夜间，内蒙古地区、西南大部、黄淮东部等地有中雨”。

现实中业务服务人员的个人经验都不相同，故预报地理信息的顺序也有可能不相同，而系统中由于每次提取的地理数据先后顺序不同，可能导致公报信息中的地理信息排列先后也不同，上段的例子就可能变成“今天白天到夜间，西南大部、内蒙古地区、黄淮东部等地有中雨”。这也不符合气象科技服务精细化的要求。通过研究降水类气象公报以及与气象中心工作人员交流，确定以下预报规则。

在预报降水气象公报时，预报顺序总体方向大致是先由西向东，再由北向南。地理一级区划的预报顺序：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。在现实的气象状况中，西北地区的形状很特殊，降水云层几乎不可能覆盖整个西北地区，所以气象公报中在预报西北地区时往往把二级地理区划甚至是行政省份作为起点。行政省份区划的预报顺序如图4-4所示；



图4-4 行政省份预报顺序图

现实中经常存在不同等级区划的地理区域发生同一天气现象，在这种情况下，按照一、二、三、四级区划的顺序来预报。但是如果一级区划的区域和其包含的二级区划的区域同时出现，那么将不在预报这个二级区划的区域。依次类推，三四级区划的关系也是这样。例如，内蒙古属于一级区划的地区，内蒙古东部属于二级区划的区域，如果出现了“内蒙古局部、内蒙古大部以及内蒙古地区”的任意一种情况，那么就不再预报类似于“内蒙古西部”之类的二级区划的内容。

预报区域范围的优先级要大于地理区划顺序。例如，青海属于三级区划的区域，内蒙古属于一级区划的区域，在预报时，应该先预报青海省，再预报内蒙古地区，因为青海地区隶属于西北地区，而西北地区的预报顺序优先级比内蒙古地区高。气象公报生成时本着符合自然语言习惯的原则，可能会对某些地区的名称进行简化预报，例如“新疆维吾尔自治区有沙尘暴、广西壮族自治区有小雨”可以简化为 “新疆有沙尘暴、广西有小雨”。

由以上规则可以得出，“今天白天到夜间，内蒙古地区、西南大部、黄淮东部等地有中雨”需要转化为“今天白天到夜间，内蒙古地区、黄淮东部、西南大部等地有中雨”。

在实际的气象预报中，雨与雪都属于降水类型，在气温0度以上，降水都为雨，反之，大多为雪。国家气象中心对降水量级由大到小分类为：1）零度以上，特大暴雨、大暴雨、大雨、中雨、小雨；2）零度以下，特大暴雪、大暴雪、暴雪、大雪、中雪以及小雪。综合的判断降水类型，选择相应句子模版，即可生成降水公报，如下所示。

17日12时至18日12时，新疆西部和北部、青海南部、甘肃中南部、西藏东部和北部、江淮西部、江南、华南、四川局部、重庆大部、贵州和云南东南部和东北部等地有小到中雨，其中，湖南南部、江西南部、福建西北部和东北部、广东局部和广西部分地区等地有大雨，其中，广东西北部和广西局部等地有暴雨。

地理区域名称需要简化以及合并，气象要素亦是如此。当Item里面的气象信息是“风”时，需要同时对风向、级别聚类。气象专家指出，相同地区出现两种风向截然相反的概率基本为0。因而当一个地方出现两种风向时，系统将对两种风向对应的角度取平均值，然后得出合并后风向的角度。例如，公报“内蒙古有3-5级东风、内蒙古4-6级东北风”，经计算，平均值落入东风的范围，故可以简化为“内蒙古有3-5级东风，内蒙古有4-6级东风”。即如果同一个地方出现两个级别，那么进行合并处理，例如，公报“新疆东部有3-5级东风、新疆东部有4-6级东风”则可以简化为“新疆东部有3-6级东风”。经过风向、级别聚类后，实验得到很多组的不同风向地理区域的数据，对所有的数据进行遍历，如果同一类中含有多个二级区划，那么进行合并处理。例如，“新疆东部有3-5级东风，新疆北部有4-6级东风”，应该化简为“新疆局部地区有3-6级东风”。

中国地大物博，气象现象种类繁多，不少地区经常发生强对流天气。通过对历史服务文本的分析，发现在某一天只会有一种强对流天气，并且强对流天气的出现往往伴随着强烈的雨雪天气，经过归纳这种气象的特征，建立如下句子模版：

1. [强对流天气] = 上述部分地区伴有短时雷雨大风、冰雹等强对流天气。
2. [强对流天气] = 上述地区局部有短时雷雨大风等强对流天气。
3. [强对流天气] = 上述地区局部伴有大风、强降雨等强对流天气。

如下是一个伴随雨雪天气的强对流天气预报例子：

27日10时至28日10时，西北西部、东北大部、华北北部和黄淮东部等地有小到中雪（雨），其中，新疆北部等地有大雪，局部地区有暴雪。四川南部、重庆西部、贵州部分地区和云南部分地区等地有小雨。渤海、渤海海峡、黄海和东海有7~8级的风。**上述部分地区并伴有短时雷雨大风或冰雹等强对流天气。**

### 4.3.3 篇章规划

经过短语规划和句子模块规划，已经可以得到全部描述气象信息的句子，在篇章规则库的基础上，与句子规划填槽类似，把每个模版的参数换成实际的句子，最终生成气象公报文本。

实验建立了“气象公报”模版，其中“[YRAR]”、“[MONTH]”、“[DAY]”、“[HOUR]”、“[AUTHOR]”、“[FAX]”和“[PHONE]”等都是公报的附加信息变量，“[WEATHER\_SUM]”则是具体存储公报内容的变量，系统可以通过修改配置文件变量的相对位置，来调整公报的格式。

## 4.4 图像与文本交互

国家气候中心投资研发了气候信息交互显示与分析系统，这个系统的主要功能是读取、解析气象数据，将气象图形与基础地理信息叠加进行显示，并提供分析处理算法、用户交互产品制作等功能，在交互产品制作完成以后，将产品保存成所需格式的图片或文件。

如图4-5为系统的用户界面，在图上任意圈出某一地方的气象观测站，即可实时的得到这些观测站的数据，并通过实验系统进行空间分析、处理与文本生成，得到相应地区的气象天气状况，进而实现了公报文本与气象云图交互。

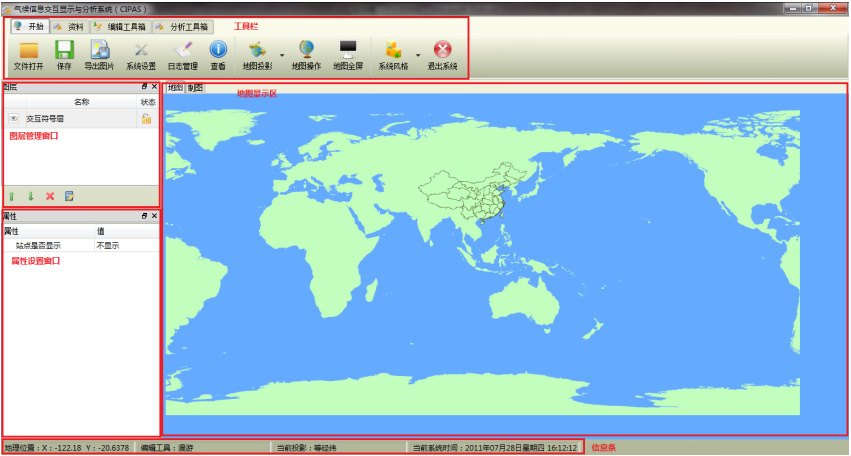


图4-5 气候信息交互显示与分析系统

# 第五章 原型系统与实验验证

在国家气象中心业务发展办公室的支持下，研究以气候信息交互显示与分析系统为平台，开发了气象公报自动生成系统。在平台上通过手工录入气象落区，如图5-1所示，进而得到气象观测站的实时气象数据，系统直接转换分析生成标准格式的气象数据，并最终形成描述指定落区的气象公报文本。

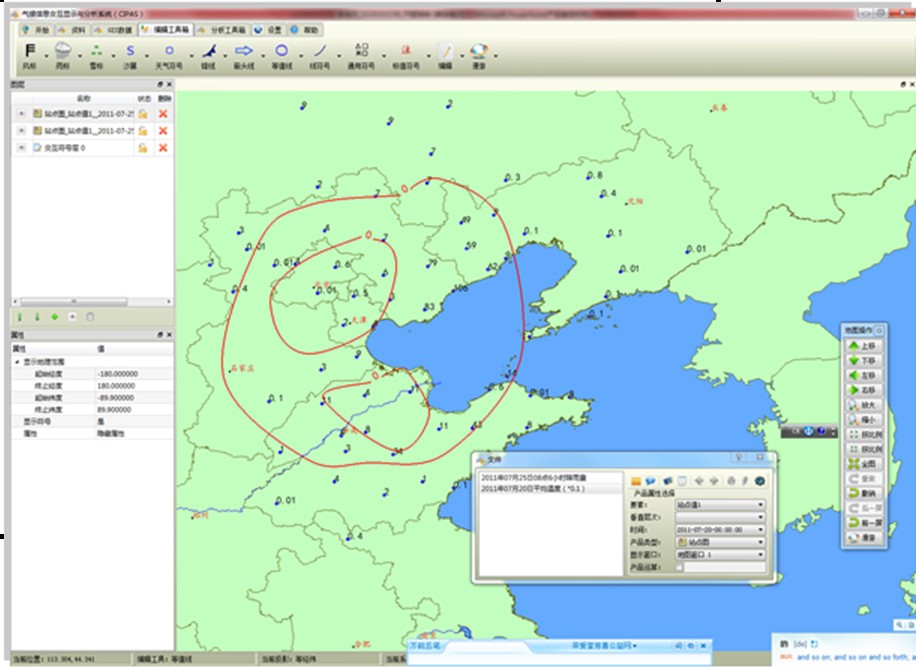


图5-1 落区选择

系统可以支持气象公报以及海洋公报的自动化生成，并能通过修改配置文件来进行阈值、结构以及语法的调整。如图5-2为系统生成全国实时的降水公报实例。TITLE的内容可以手动设定，例如气象公报的内容为干旱时，可以添加抗旱的相关指示等。

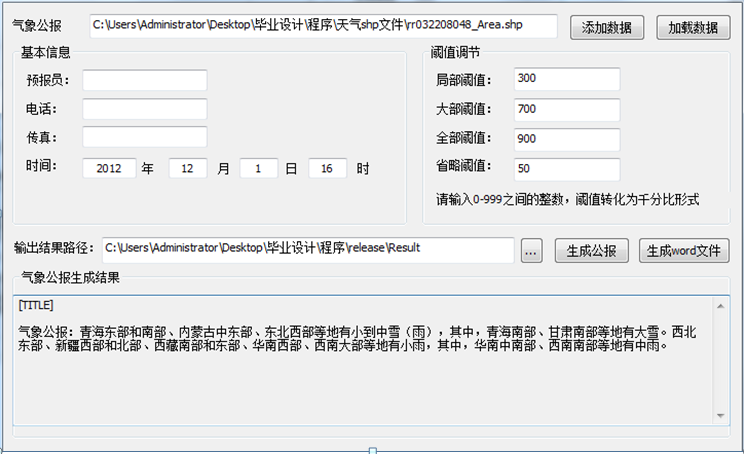


图5-2 系统生成降水公报实例

系统自动生成的海洋公报如下所示：1日20时至17日20时，北部湾有5-6级的东风。琼州海峡、南海有7-8级的东风。台湾海峡东部有8-9级的东南风。东海南部有6-10级的北风。巴士海峡有7-10级的东北风。台湾海峡有7-10级的东南风。其他海域天气海况条件较好。

在已提取的气象服务文本特征和成熟的GIS空间分析方法的基础上，系统已经涵盖强对流、雨雪、风这些气象领域，并在以后的研究中对高温、沙尘、霜冻等领域进行扩展。

在此，实验随机选择了某天气象中心发布的GIS数据，经过系统空间分析、处理，进而生成气象公报。实验将此公报与原始的数据进行分析对比，并与业务员人工手写的公报进行比较，验证了系统能如实的描述当日的气象信息，能代替业务员手动撰写。

业务员手动撰写的气象公报如下：1日17时至2日17时，青海东南部、内蒙古中部和东北西部等地有小到中雪，甘肃南部等地有大雪。西北局部、新疆大部、西藏南部、华南西部以及西南大部有小雨，其中，华南、西南局部地区有中雨。新疆北部有4-5级东风，东北大部有4-5级的西南风，内蒙古东部有4-6级的西北风，内蒙古西部有5~6级的东北风，新疆东部有6~7级的东风。

经系统处理原始GIS数据后分析生成的公报文本如下：1日17时至2日17时，青海东部和南部、内蒙古中东部和东北西部等地有小到中雪（雨），其中，青海南部和甘肃南部等地有大雪。西北东部、新疆西部和北部、西藏南部和东部、华南西部和西南大部等地有小雨，其中，华南中南部和西南南部等地有中雨。新疆北部有4~5级的东风。东北北部有4~5级的西南风。内蒙古东部有4~6级的西北风。内蒙古西部有5~6级的东北风。新疆东部有6~7级的东风。

通过随机的对比分析，可以看出，人工撰写的公报与系统自动生成的公报内容大致相同，而且系统生成的公报内容更为详细，基本涵盖了所有的气象信息点，且用语准确，也避免了不同业务员撰写风格差异的问题。

实验随机抽取了100篇业务员撰写的气象公报以及对应的GIS数据，并使用系统对这些数据分析、计算，并生成相应的气象公报。本人制订了一套5级打分的方法来对这两组气象公报进行人工的评判打分，分数与相似度正相关，5分为非常好，4分为好，3分为中等，2分为不好，1分为极差，其中1分、2分算不及格。经过统计，结果如表5-1所示。

表5-1 公报打分结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 | 4 | 3 | 2 | 1 | 合格率 |
| 21 | 49 | 26 | 3 | 1 | 96% |

由表可以看出，系统生成的气象公报能够比较好的描述全国范围以及海域的气象信息，本文提出的气象文本生成与交互方法也有足够的可行性。

# 第六章 总结与展望

气象部门的业务人员通过对获取或者生成大量的气象观测资料、模式预报资料等的人脑解释综合分析和描述，从而产生面向不同服务对象的大量气象服务文本信息，这些文本信息规模庞大，不同类型的文本风格特征迥异，但面对的人群却十分广泛，对公众的影响特别显著，是气象服务业务发展水平的一个体现。本文首先运用了自然语言统计学研究方法，选取了气象部门（天气和气候领域）人工撰写的典型文本作为研究对象，对“天气预报信息”、“重要气候信息”和“每日天气提示”中的词和短语进行了特征抽取和综合分析，提取了文本中地理区划、方向、范围、气象要素等用户特征。这些特征为研究后续的文本生成与交互为技术上的可行性提供了保证。

气象的精准性和复杂性之间的矛盾非常突出，如何能实时的、准确的、快速的对气象状态进行描述，同时又能生成符合公众阅读习惯的文本并与图像交互，决定了与众不同的数据分析以及文本生成方法。如何对大量复杂的GIS数据进行空间分析并提取出标准格式的气象数据成为了关键性问题。本文建立了提取气象数据的模型，介绍了空间分析的详细步骤，并解决了对GIS数据实时的分析、处理与转换的难题。研究最后建立了基于自然语言的文本生成模型，并实现了气象公报的自动化生成。实验研发的系统以国家气候信息交互显示与分析系统为平台，由业务员在平台上手工录入气象落区，系统自动读入气象观测站的实时气象数据，进而生成指定落区的公报，实现了文本与图像的交互。

本文的创新之处有两点：一是气象服务文本的特征提取与用语分析，填补了国内气象领域特征提取的空白，为气象文本的自动化生成奠定了技术基础，也对对业务人员尤其是新预报员的人工撰写有较大的指导意义。二是气象服务文本的生成与交互，研究以国家气象服务平台为依托，实现了实时的气象数据（与文本的交互，提高了气象文本自动化生成与交互的水平。

然而气象服务产品种类繁多，系统尚不能很好的满足公众大部分的需求，在GIS空间分析领域，仍有许多高效便捷的空间分析方法需要探讨研究，在文本的自动化、半自动化生成与交互领域，更有非常大的研究空间和潜力值得本人去学习与探索。

# 参考文献

1. 中国气象局业务技术体制改革总体方案[R]. 中国气象局办公室文件[EB/OL]. [2006-01-12].http://www.cma.gov.cn/qxzt/yjtzgg/wjhb/zt/t20060406\_170222.phtml.
2. 李学晋.模糊语言在天气预报中的应用[J].时代文学,2009(4):16-129.
3. 王耀生.天气过程语言[J].气象,1990(5):49-54.
4. Kelly Potts. Word’s first forecast technology aimed at reducing winter road accidents, the Communications Team, Office of External Affairs, University of Aberdeen,King’sCollege,Aberdeen.[EB/OL].http://www.abdn.ac.uk/news/archive-details-4153.php.
5. 郑国光.开拓创新，进一步推动气象事业实现更大的发展[R].2010全国气象局局长会议工作报告,2010,7-8,北京.
6. 江铭虎.自然语言处理[M].高等教育出版社,2006.12,北京.
7. 宗成庆.统计自然语言处理[M].清华大学出版社,2008.5,北京.
8. 刘群,张华平,骆卫华,孙健.自然语言理解(译著)[M].电子工业出版社,2005.1,北京.
9. Ehud Reiter, Robert Dale. Building natural language generation system [M], Cambridge, UK: Cambridge University Press. ISBN 0-521-02451-X.
10. Hua-Ping ZHANG, Qun LIU, Hong-Kui YU, Xue-Qi CHENG, Shuo BAI. Chinese Name Entity Recognition Using Role Model. Special issue "Word Formation and Chinese Language processing" of the International Journal of Computational Linguistics and Chinese Language Processing, vol.8, No.2, 2003, pp. 29-602.
11. 吴焕萍,罗兵,王维国等.GIS支持下的决策气象服务系统建设研究[J].应用气象学报,2008.19(3):380-383.
12. E. Goldberg, N. Driedger, and R. Kittredge. Using Natural-Language Processing to Produce Weather Reports[J]. IEEE Expert: Intelligent Systems and Their Applications, Volume 9, Issue 2 (April 1994) pp:45-53.
13. Eli Goldberg. FOG: Synthesizing Forecast Text Directly from Weather Maps[R]. Artificial Intelligence for Applications, 1993. Proceedings of Ninth Conference, Orlando,FL,USA. Available at: www.aclweb.org/anthology/C/C86/C86-1132.pdf.
14. KiRredge Richard, d Lavoie. Benoit Meteocogent：A knowledge-based Tool for Generating Weather Forecast Texts. In Proceedings of the American Meteorological Society AI Conference(AMS-98)[C], 1998, Phoenix, Arizona.
15. Ross Turner, Somayajulu Sripada, Ehud Reiter, Ian P Davy. Building a Parallel Spatio-Temporal Data-Text Corpus for Summary Generation[C]. In Applications and Innovations in Intelligent Systems (In Proceedings of the LREC2008 Workshop on Methodologies and Resources for Processing Spatial Language, Marrakech, Morocco, 31 May 2008) XV: pages 75-88.
16. Turner R., Sripada S., Reiter E. and Davy I. Using Spatial Reference Frames to Generate Grounded Textual Summaries of Georeferenced Data In Proceedings of INLG08, Salt Fork, Ohio, USA, 12-14th June 2008.
17. Somayajulu G Sfipada, Ehud Reiter, Ian Davy, Kristian Nilssen．Lessons from deploying NLG technology for marine weather forecast text generation．Proceedings of PAIS[C]．2004, 2004.
18. 曾庆辉, 姚天畴, 顾建峰. 天气预报文本生成系统中的天气图处理器[J]．计算机工程, 2000, 26(3)：28—30.
19. 熊安元.“数字气象”发展的战略分析[J].地球信息科学,2002(12):101-104.
20. 姚天昉,黄小戎,张冬茉,王纤. 天气预报汉语文本生成模型的研究[J]. 江西师范大学学报（自然科学版）. 1998(10): 102-111.
21. 姚天昉,汤学彦.文本规划中焦点移动控制算法的研究[J].软件学报,2000, 11 (2) : 277- 284.
22. 张冬茉,葛永,姚天昉. 多语种自然语言生成系统中的预映射句子规划器[J]. 计算机研究与发展,2001(4):467-474.
23. Deerwester S, Dumais, S.T.Furnas, G.W, Landauer,T.K, &Harshman,R.(1990) Indexing by Latent Semantic Analysis, Journal of the American Society For Information Science,41(6),391-407．
24. 冯志伟.自然语言生成系统的建造[M].北京:北京大学出版社，2010年．
25. Abney Steven, Parsing by Chunks, In: Robert Berwick, Steven Abney and Carol Tenny(eds.), Principle-Based Parsing, Kluwer Academic Publishers, 1991, pp.257-278.
26. 李素建，汉语组块计算的若干研究，中国科学院计算技术研究所博士学位论文, 2002.
27. 中国气象局预测减灾司,中国气象局国家气象中心.中国气象地理区划手册[M],气象出版社,2006.8,北京.
28. L．A．Zadeh, A computational approach to fuzzy quantifiers in natural language，Computers and Mathematics with Applications，9(1983)149-184．
29. 冯祥胜.基于Sugeno积分语言量词模型的气象文本自动生成技术研究[J]. 科技广场,2009(09):36-38.
30. Matheron B. The theory of Regionalized variables[M].Fontainebleau:Centre de Morphologie Mathematique,1971.1-200.
31. 冯克忠,万庆.基于组件技术的GIS广义空间分析[J].地球信息科学,2003,5(1):62-66.
32. VIJAYKHATRI, SUDHA RAM, RICHARDT SNODGRASS. STUSM: Bridgingthe semantic gap with a spat io-temporal conceptual model[R] . A TIME CENTER TR-64, 2001.
33. ERWIG M, SCHNEIDER M. Spat io-temporal predicates IEEE trans on knowledge and data [J] .Engineering, 2002, 14(4) : 881- 901.
34. EGENHOFER M J. Reasoning about binary topological relations[R]. Advances in Spatial Databases: Springer Verlag, 1991.
35. Hua-Ping Zhang, Huan-Ping Wu, Jian Gao. Meteorological Bulletin Automatic Generation based on Spatio-Temporal Reasoning. Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (ICMLC 2011)[C]. p 1927-31, 2011.
36. Hua-Ping ZHANG, Jian GAO, Qian MO,He-Yan HUANG. Incorporating New Words Detection with Chinese Word Segmentation. In Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010) [C]. Beijing, China.2010.8 .p249-251.
37. Lavoie, B., and O. Rambow. A Framework for Customizable Generation of Multi-Modal Presentations. To appear in Proceedings of the 36th Meeting of the Association for Computational Linguistics (ACL ’98), Montréal, Canada.
38. Ross Turner, Yaji Sripada, Ehud Reiter. Generating Approximate Geographic Descriptions[C]. Proceeding of the 12th European Workshop on Natural Language Generation. Athens, Greece, 2009: 42-49.
39. Zhang XF, Cui WH. Spatio-Temporal analysis and modeling based on the integration of GIS and CA model. Journal of Image and Graphics, 2000, 5A(12):1012~1018 (in Chinese with English abstract).
40. Roger Evans, Paul Piwek, and Lynne Cahill. 2002. What is NLG? In Proceedings of the Second International Conference on Natural Language Generation, Pages 144-151.

# 攻读学位期间发表论文与研究成果清单

1. 李笑侃.基于全生命周期的云计算安全隐患分析模型.VARA.2011.
2. 李笑侃.典型云计算系统模拟仿真平台的设计与实施.VARA2012.

# 致谢

时光飞逝，岁月如梭，如歌。两年半的研究生美好求学生活即将结束，站在毕业的门槛上，往日的奋斗、拼搏已成为丝丝的记忆。北京理工大学以优良的学风、严谨的科研氛围不断地培育着我，“德以明理，学以精工”的校训也深深的刻在我的心里。在毕业论文即将完成之际，我谨向所有关心、帮助、爱护我的人们表示最诚挚的感谢和最美好的祝福。

本论文从选题到完成，每一步都是在李玉岗、张华平两位老师的悉心指导下完成的，李老师和张老师以诲人不倦的高尚师德，精益求精的工作作风，严谨的治学态度，丰富的实践经验、对学科发展方向的敏锐眼光给予了我极大的启迪和引导。两位老师对我的工作提出了严格的要求，也促使我对研究的方法、精神、规律有所领会，这些收获也是今后工作以及为人的重要基础。在此，我向两位老师表示深深的感谢祝福！

当然，本论文的完成也离不开其他的各位老师、同学和朋友的关心与帮助。在此也要感谢国家气象中心的吴焕萍博士，北京理工大学经管学院的赵燕平老师对我的论文开题、初稿、预答辩期间提出的宝贵意见，还要感谢实验室的同门师兄师妹们，在科研过程中给我以许多鼓励和帮助。

另外感谢父母在我求学生涯给与我无微不至的关心与照顾，同时，还要感谢梁超、李胜、李玉鹏对我的爱护、包容以及帮助，愿友谊长存！

李笑侃

2012.12.2