



基于时空推理的气象公报自动生成

Meteorologic Bulletin Automatic Report Generation based on Spatio-Temporal Reasoning

张华平 博士

副教授 研究生导师

北京理工大学计算机语言处理研究所 副所长

<http://hi.baidu.com/drkevinzhang/>

<http://t.sina.com.cn/drkevinzhang/>



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



纲要

- 研究背景
- 相关工作
- 研究目标
- 时空推理模型
- 气象公报生成的实验效果



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



研究背景

- 气象公报文本信息是最适合大众接受的气象服务形式。
- 气象公报语言千变万化，编写准确合理的气象服务文本需要较强的气象专业背景与实际工作经验，但是其中存在大量机械重复的工作。
- 气象公报是“图文并茂”的，但是其文本信息的自动化生成技术离实际需求甚远。





研究现状

➤ 早期的基于模板的文字替换（Template-based Computer Worded Forecasting, CWF）技术

方法:首先对要生成的气象信息文本建立一个模板，与气象信息有关的文字设置为模板的槽。将实际的气象信息转换为合适的文字填入模板的槽中，文本就生成了。

CWF技术的显著优点是简单，容易实现；

缺点是需要大量的模板，由此带来了管理上的困难，且不能进行有效的数据分析，同时还不能形成像自然语言那样流畅的文本，因而限制了应用范围。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



研究现状

➤ 基于NLG(Natural Language Generator)的加拿大FoG系统和英国的SumTime_Meteo系统

FoG系统主要用于生成海洋天气预报文本，数据经过简单分析，然后将被选择的内容发送到语言处理器，其内容被结构化后，就形成了语言预报文本（英语和法语）。

SumTime_Meteo采用了两年的人工海洋天气预报文本资料以及相应的天气预报数据，建立了天气预报语料库，然后通过数据分析、文档规划、宏观规划以及表现来实现预报文本的计算机自动生成。





研究现状

➤ 上海交通大学的MLWFA

MLWFA设计实现天气预报文本的计算机生成，这个项目的研究重点在于自然语言生成框架内的技术研究，在生成模型、文本规划、句子优化、文本连贯性表达等方面做了较为深入的研究。

上海交通大学于1999年开始研究，而最近十年内没有检索到其它的相关研究成果。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



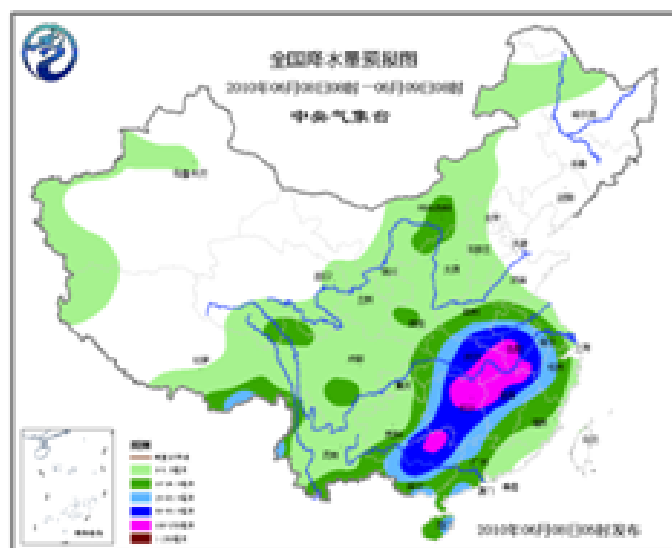
研究现状

- 面向气象服务文本生成的多维时空气象数据的信息提取与推理方法还没有系统、深入的研究，无法形成文本生成的基础能力。
- 国外基于自然语言生成的、面向天气预报的文本生成方法已经得到了初步的研究；而在国内还没有引起足够的重视，除了一些业务单位采用简单的CWF技术外，最近的研究是十多年前上海交通大学有一些初步的研究，但并没有形成有效的应用，也没能得到持续深入的进一步研究。



未来三天，青藏高原中东部及其以东大部地区自西向东、自北向南将有一次明显降雨过程，其中，湖北大部、河南南部、安徽、江苏西南部、浙江西部、福建大部、江西、湖南、广西、广东中西部、海南等地有大到暴雨，局地有大暴雨。另外，西藏东南部、云南西北部 and 东南部的部分地区有大雨。上述部分地区并伴有短时雷雨大风或冰雹等强对流天气。

具体预报如下：



8日08时至9日08时，新疆西部、西藏东部和西部的部分地区有小到中雨（雪）或雨夹雪，西北地区中东部、内蒙古中西部和东北部、黑龙江西部、山西、河北西部、黄淮大部、江淮、江汉、江南、华南大部以及西南地区

等地有小到中雨或阵雨，其中，**西藏东南部、云南西北部、河南南部、湖北大部、安徽大部、江苏西南部、浙江西部、江西中北部、湖南大部、贵州东南部、广西大部、广东西部、海南东部等地有大到暴雨**，其中，**湖南东北部、江西北部、湖北东部、安徽西南部、广西东北部等地的部分地区有大暴雨**；上述部分地区还将伴有**短时雷雨大风或冰雹等强对流天气**。新疆西部和北部、甘肃西部、宁夏、内蒙古中西部等地有**4~6级风**。北部湾有**5~6级偏南风**。



研究目标

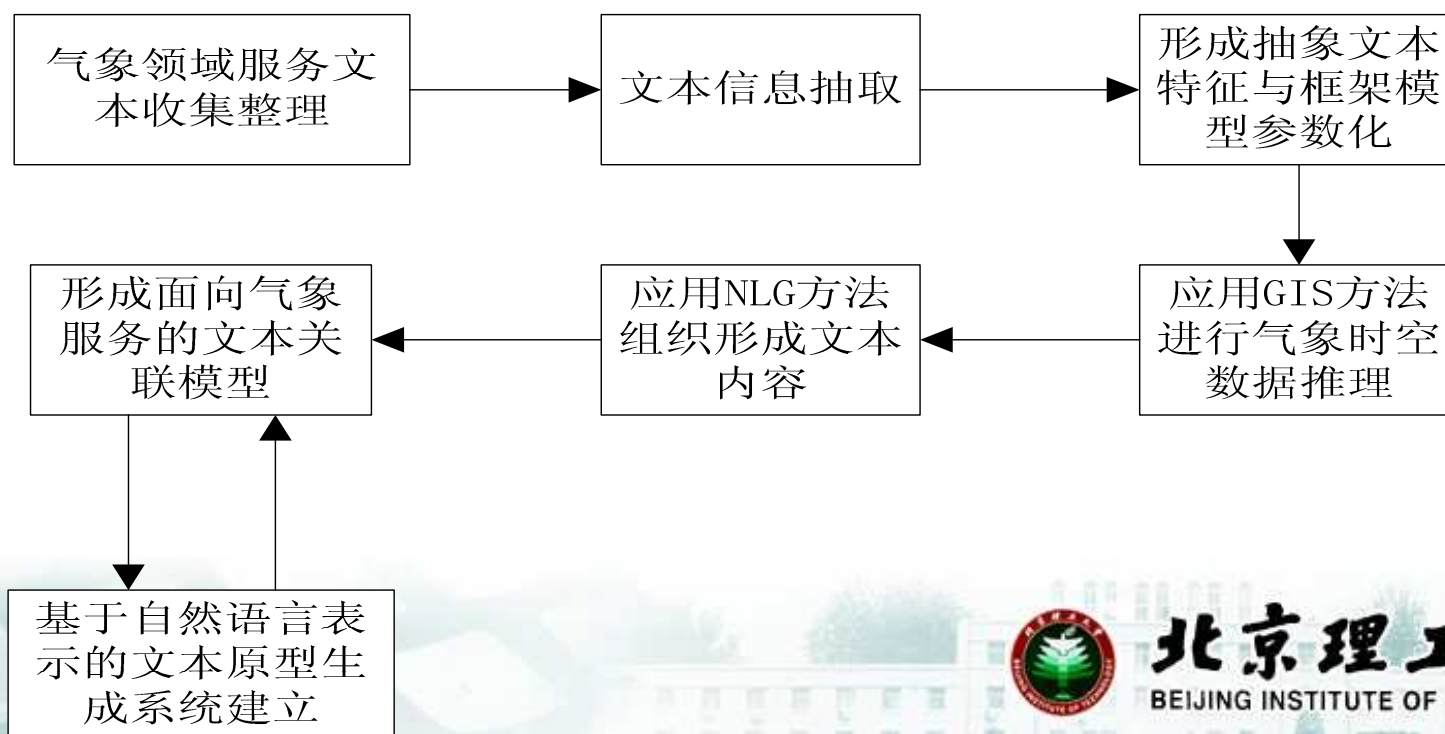
- 研究气象公报文本语言组织的特点，提取气象文本特征并建模，建立起基于计算语言的气象语言体系。
- 设计一个气象公报自动生成系统，将气象云图转换成计算机能识别的信息，经过一系列的计算和规则配置，生成易于大众理解和符合预报员写作习惯的公报文本。
- 预报员每天撰写气象公报的时间将会节省下来，提高工作效率。





技术路线

- 时空推理是指对占据空间并随时间变化的对象所进行的推理。
- NLP (Natural Language Processing) 是人工智能（AI）的一个子领域。





创新之处

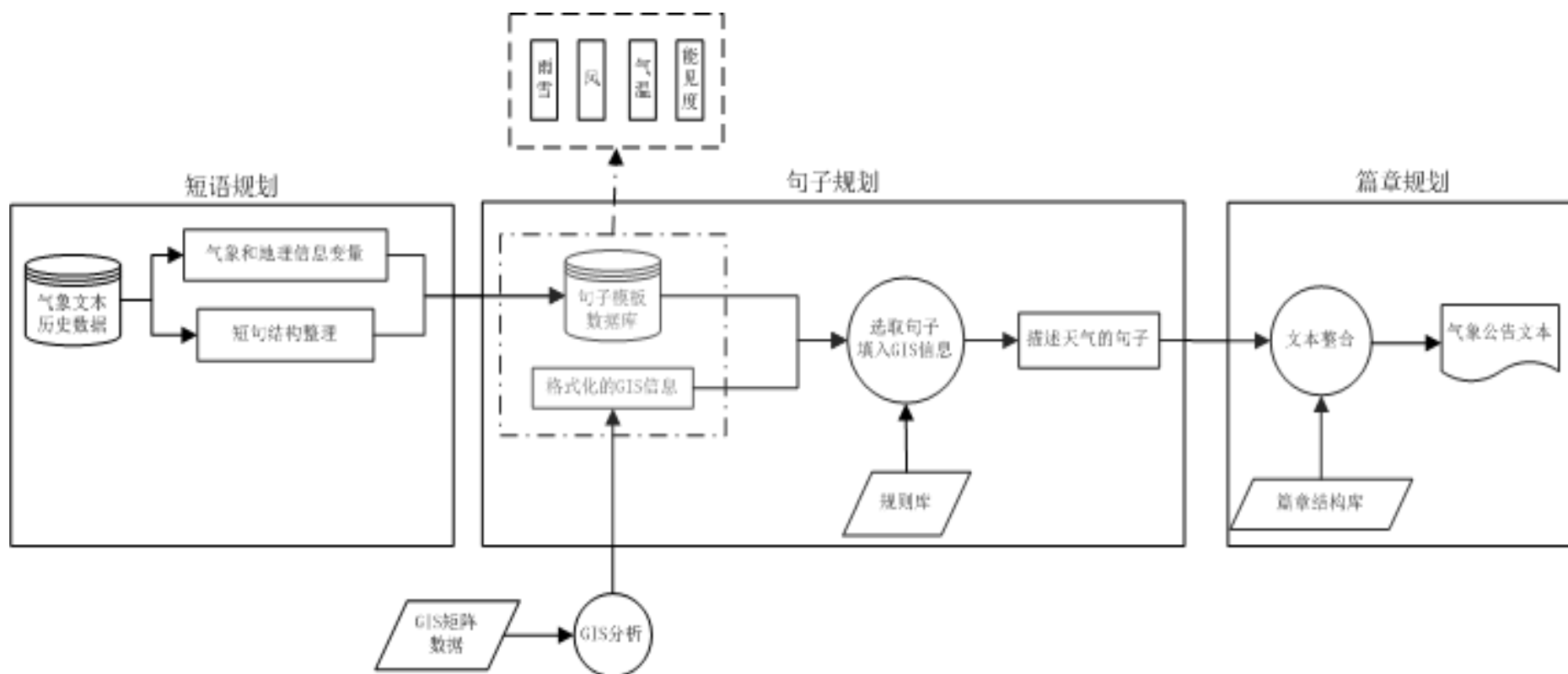
- 学科交叉：文本挖掘、自然语言理解、气象
- 建立一套实用的时空数据推理算法，以分析适于语言理解的气象描述
- 实际需求驱动



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



系统设计框架





气象原始数据

2010 6 7 17 0

LINES: 0

LINES_SYMBOL: 0

SYMBOLS: 0

CLOSED_CONTOURS: 12

4 113

122.585	53.649	0.000	122.404	53.608	0.000	122.224	53.567	0.000	122.045	53.525	0.000
121.866	53.482	0.000	121.688	53.439	0.000	121.510	53.394	0.000	121.334	53.348	0.000
121.158	53.300	0.000	120.983	53.251	0.000	120.810	53.200	0.000	120.638	53.147	0.000
120.467	53.092	0.000	120.298	53.034	0.000	120.130	52.974	0.000	119.964	52.911	0.000
119.800	52.845	0.000	119.637	52.776	0.000	119.477	52.704	0.000	119.319	52.628	0.000
119.163	52.549	0.000	119.010	52.466	0.000	118.859	52.379	0.000	118.772	52.327	0.000
118.687	52.273	0.000	118.603	52.218	0.000	118.521	52.161	0.000	118.442	52.104	0.000
118.365	52.046	0.000	118.291	51.987	0.000	118.220	51.926	0.000	118.153	51.866	0.000
118.153	51.866	0.000	118.041	51.791	0.000	117.930	51.715	0.000	117.819	51.639	0.000
117.709	51.562	0.000	117.601	51.484	0.000	117.494	51.404	0.000	117.388	51.322	0.000
117.202	51.168	0.000	117.027	51.008	0.000	116.869	50.842	0.000	116.732	50.670	0.000
116.624	50.494	0.000	116.547	50.313	0.000	116.508	50.128	0.000	116.519	49.878	0.000
116.597	49.628	0.000	116.734	49.383	0.000	116.924	49.149	0.000	117.158	48.932	0.000
117.432	48.736	0.000	117.737	48.567	0.000	118.099	48.420	0.000	118.485	48.309	0.000
118.890	48.233	0.000	119.306	48.188	0.000	119.728	48.175	0.000	120.149	48.189	0.000
120.564	48.229	0.000	121.037	48.307	0.000	121.497	48.411	0.000	121.948	48.531	0.000
122.394	48.660	0.000	122.841	48.788	0.000	123.291	48.908	0.000	123.749	49.010	0.000
124.117	49.072	0.000	124.492	49.118	0.000	124.873	49.148	0.000	125.260	49.161	0.000
125.652	49.158	0.000	126.048	49.138	0.000	126.449	49.102	0.000	126.846	49.053	0.000
127.242	48.997	0.000	127.633	48.945	0.000	128.017	48.905	0.000	128.390	48.888	0.000
128.749	48.902	0.000	129.091	48.959	0.000	129.445	49.077	0.000	129.769	49.248	0.000



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



文本结构分解

(8日08时至9日08时), (新疆西部、西藏东部和西部的部分地区)有小到中雨(雪)或雨夹雪, (西北地区中东部、内蒙古中西部和东北部、黑龙江西部、山西、河北西部、黄淮大部、江淮、江汉、江南、华南大部以及西南地区等地)有小到中雨或阵雨, 其中, (西藏东南部、云南西北部、河南南部、湖北大部、安徽大部、江苏西南部、浙江西部、江西中北部、湖南大部、贵州东南部、广西大部、广东西部、海南东部等地)有大到暴雨, 其中, (湖南东北部、江西北部、湖北东部、安徽西南部、广西东北部等地)的部分地区有大暴雨; 上述部分地区还将伴有短时雷雨大风或冰雹等强对流天气。 || (新疆西部和北部、甘肃西部、宁夏、内蒙古中西部等地) 有4~6级风。(北部湾) 有5~6级偏南风。





气象GIS接口

趋势信息

```
<item>
  <from>
    <day_id>1</day_id>
    <basic_id>1</basic_id>
  </from>
  <to>
    <day_id>2</day_id>
    <basic_id>16</basic_id>
  </to>
  <orientation>SouthEast</orientation>
</item>
<item>
  <from>
    <day_id>1</day_id>
    <basic_id>1</basic_id>
  </from>
  <to>
    <day_id>2</day_id>
    <basic_id>17</basic_id>
  </to>
  <orientation>NorthEast</orientation>
</item>
```

趋势中的降水圈信息

```
<item>
  <id>43</id>
  <day_id>3</day_id>
  <type>雨雪</type>
  <weather_name>大暴雨</weather_name>
  <value>100</value>
  <area>29482</area>
</item>
```

每日降水圈信息

```
<item>
  <id>7</id>
  <geo_name>西北地区</geo_name>
  <proportion>0.010</proportion>
  <area>29176</area>
  <type>雨雪</type>
  <weather_name>暴雨</weather_name>
  <value>50</value>
</item>
<item>
  <id>8</id>
  <geo_name>华北地区</geo_name>
  <proportion>0.406</proportion>
  <area>175486</area>
  <type>雨雪</type>
  <weather_name>小雨</weather_name>
  <value>0</value>
</item>
<item>
  <id>9</id>
  <geo_name>华北地区</geo_name>
  <proportion>0.047</proportion>
  <area>20530</area>
  <type>雨雪</type>
  <weather_name>中雨</weather_name>
  <value>10</value>
</item>
```

降水圈包含关系信息

```
<item>
  <id_big>2</id_big>
  <id_small>34</id_small>
</item>
<item>
  <id_big>3</id_big>
  <id_small>35</id_small>
</item>
<item>
  <id_big>8</id_big>
  <id_small>36</id_small>
</item>
<item>
  <id_big>11</id_big>
  <id_small>37</id_small>
</item>
```



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



气象语言特征分析

- 训练数据：2008~2009两年内所有的气象公报文本，总共1963个公报文件。
- 数据抽取：气象专业名词变量，地理区划名词变量，地理区划描述顺序规则
- 数据分析：词频统计、二元邻接词统计





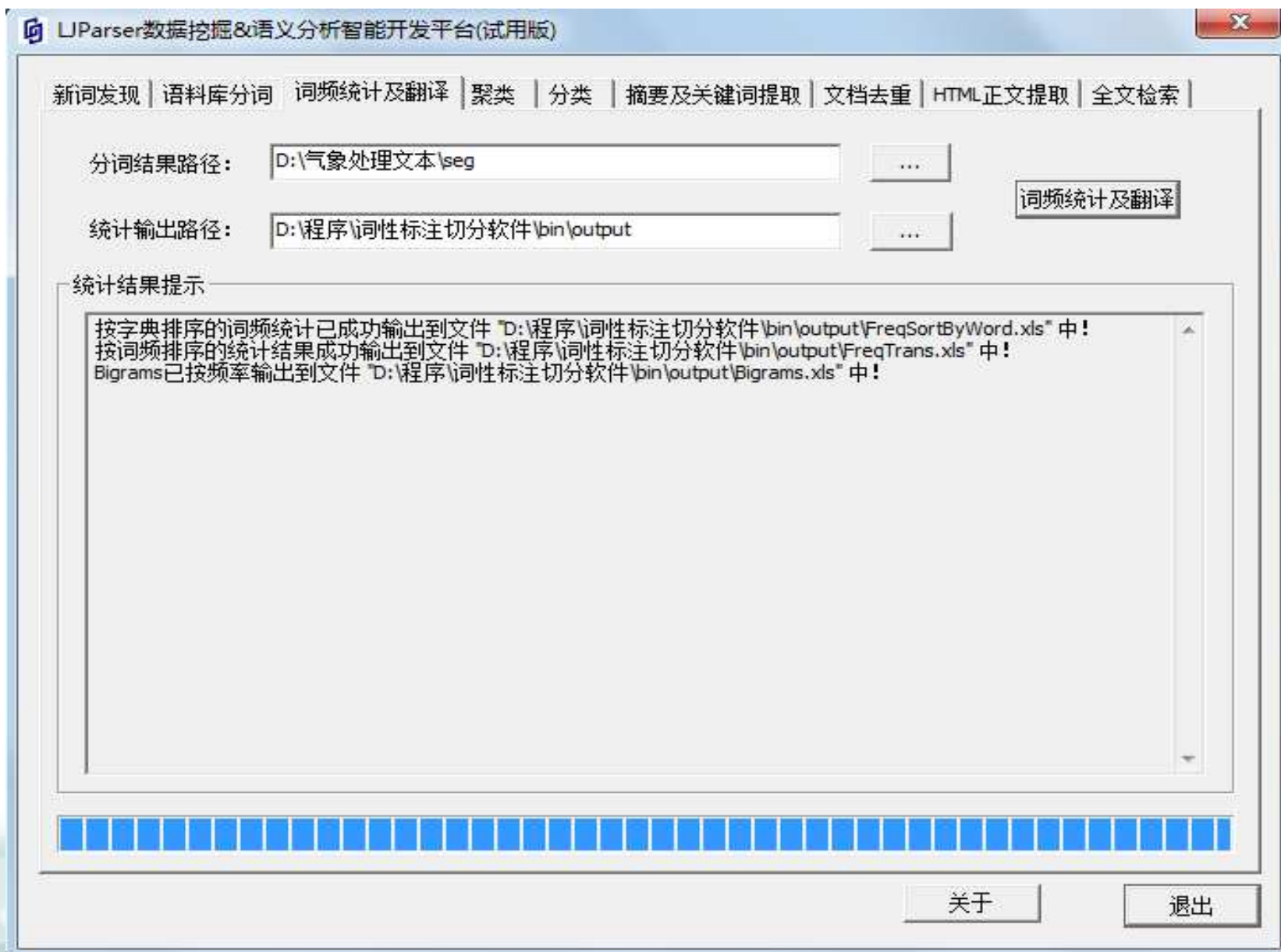
气象语言特征分析

- 气象名词：中雨、阵雨、暴雨、雨夹雪、冰雹、扬沙、浮尘、雾……
- 地理名词：西北地区、黄淮地区、西南地区、新疆、内蒙古、山东……
- 方位名词：东部、西部、南部、北部、东北部、西北部、东南部、西南部
- **以上名词变量特点：数量有限、指代明确**





气象语言特征分析





气象语言特征分析

二元词组邻接分析

二元词对总数为：17314			
前一个词	后一个词	共现频次	转移概率
新疆	山口	580	0.059371
黑龙江	西北部	579	0.160343
四川盆地	、	578	0.216317
黄淮	西部	577	0.107269
东北地区	、	576	0.145565
到	明天	573	0.273378
西部	的部分地区	571	0.044718
大风	。	565	0.102746
黄海	大部	565	0.175411
度	，	564	0.595565
黄淮	大部	564	0.104852
：	今天	561	0.076818
地	的	559	0.0267
贵州	大部	552	0.11582
广西	南部	548	0.135778
东南部	等	546	0.073934

词频统计

总词数为：2468，所有词的平均频率为：486.998784				
词语	词频	一元概率		
雷暴	541	0.00045		
雷电	68	0.000057		
雷雨	3684	0.003065		
雷阵雨	1719	0.00143		
雷州	231	0.000192		
雷州市	1	0.000001		
冷	4	0.000003		
冷锋	4	0.000003		
冷空气	2881	0.002397		
冷暖	56	0.000047		
冷暖空气	1235	0.001028		
冷涡	132	0.00011		



气象公报模板建立

- 1. 所有描述气象的语言只包含4个领域：雨、雪、风、气温、能见度，针对4个领域分别建立相应的气象模板。
- 2. 短语模板—>句子模板—>段落模板—>篇章模板





气象公报模板建立

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[因素]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]



气象公报模板建立

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{, [地点]的风力可达[级数]}。
[地点]风力[级数]级、瞬时阵风可达[级数]级。

雨雪 = [地点]{将}有[雨雪强度]{, 其中, [地点]有[雨雪强度]}{, 局地有[雨雪强度]。}{, 上述地区局部并伴有短时雷雨大风或冰雹等强对流天气}。

气温 = {时间}[地点]气温将[升降][度数]℃{, 其中[地区][升降]温幅度可达[度数]℃}。
{时间}[地点]{以[天气]为主, }气温逐渐回升。
{时间}[地点]气温将下降[度数]℃, {[地点]}局地降温幅度可达[度数]℃以上。
{时间}[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 = {时间}[地点]{将}有[大气透明度]{, [地点]能见度不足[能见距离]米}。
{时间}[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。
{时间}[地点]有[大气透明度]{, 其中, [地点]有能见度小于[能见距离]的雾}{, [地点]能见度不足[能见距离]米}。
{时间}[地点]有[大气透明度]天气、部分地区有[大气透明度]。

时间 = 今天白天到夜间,
今天下午到夜间,
今天白天,
今天夜间,
今天早晨,
今天早晨到上午,
[DAY]日开始,
[DAY]日起,
[DAY]日早晨和夜间,



气象变量分析

➤ 地理变量

一级气象地理区域↻	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↻
二级气象地理区域↻	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↻
三级气象地理区域↻	各行政省份或直辖市↻
四级气象地理区域↻	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。↻

➤ 方位词

大部、局部、东部、西部、南部、北部、中部、中大部、中西部、中南部、中北部、东南部、西南部、东北部、西北部

➤ 天气变量

雨雪：小雨、中雨、大雨、暴雨、大暴雨、特大暴雨

小雪、中雪、大雪、暴雪、大暴雪、特大暴雪



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



空间推理

➤ 描述顺序规则:

1. 预报的区域包括中国境内地区，预报的范围是由大到小
2. 预报某地区时一般都不会预报该地区的全称。
3. 在天气预报中预报区域范围的优先级高于预报区域顺序



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



空间推理





空间推理

➤ 地理变量包含关系:

四级 $\in$ 三级

二级 $\in$ 一级

三级 $\in$ 一级

二级和三级的关系是并列的

一级	二级	三级
内蒙古地区	内蒙古东部、内蒙古中部、内蒙古西部	
西北地区	西北地区东部、西北地区西部、西北地区中部	青海省、陕西省、甘肃省、宁夏回族自治区、新疆维吾尔自治区、
华北地区	华北东部、华北西部	山西省、河南省、河北省、山东省、北京市、天津市
东北地区	东北东部、东北西部	吉林省、辽宁省、黑龙江省
西藏地区	西藏北部、西藏东部、西藏南部、西藏西部	
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
江汉地区	江汉东部、江汉西部	湖北省、河南省
江南地区	江南东部、江南西部、江南中部	湖北省、江西省、浙江省、福建省、安徽省、江苏省、上海市、湖南省
华南地区	华南东部、华南西部、华南中部	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	河南省、江苏省、安徽省、山东省

三级	四级
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
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
安徽省	安徽省北部、安徽省南部、安徽省中部





空间推理

➤ 描述逻辑:

设西南地区降水面积与区划面积的比例为 P ，西南地区里的三级区划的比例为 p

if($P \leq 0.05$)

忽略输出

else if($P \geq 0.9$)

输出：西南

else if($0.7 \leq P < 0.9$)

输出：西南大部

else if($0.3 < P < 0.7$)//若不够0.7，进入二三层级分析

{

//判断该输出二级还是三级

if(IS_Second())//若输出二级

输出：西南东部和南部（例）

else if(IS_Third())//若输出三级

{

//逐个分析每个省的信息

if($p \leq 0.05$)

忽略输出（例）

else if($p \geq 0.9$)

输出：云南（例）

else if($0.7 \leq p < 0.9$)

输出：云南大部（例）

else if($0.3 < p < 0.7$)//若不够0.7，进入四级层级分析

输出：云南西北部|云南局部（例）

}

}



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



空间推理

➤ 描述逻辑:

二级区划描述:

建立两套叠加的二级区划方案，比如东北地区包括**东北东部**和**西部**，同时，东北地区也可以划分成**东北南部**、**中部**和**北部**。

```
if( IS_Second() )//若输出二级
{
  If(Rgionalization_1st is preferred)
  输出: 东北东部和西部
  if(Rgionalization_2st is preferred)
  输出: 东北中南部
}
```



四级描述方式同二级描述类似



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



空间推理

➤ 地理区划分类:

1. 时间
2. 雨雪种类
3. 雨雪量级（比如：小雨、中雨、大雨）和风级

➤ 实例:

25日14时至26日14时，新疆南部和西部、青海大部、甘肃部分地区、宁夏、陕西、内蒙古西部和东部、黑龙江北部和东部、山西部分地区、河南大部、安徽大部、江苏南部、江南大部、福建西北部和东北部、广东部分地区、广西、海南和西南等地有**小到中雨**，其中，青海东部、甘肃东部、陕西中南部、黄淮西部、江淮西部、安徽中南部、江汉、湖北、湖南部分地区、江西北部、四川南部和东部、重庆大部、贵州部分地区和云南西南部等地有**大到暴雨**，其中，湖北局部和江南西部等地有**大暴雨**。东海南部有**7~8级**的风。台湾海峡、台湾以东、巴士海峡、北部湾、琼州海峡和南海大部有**9~10级**的风。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

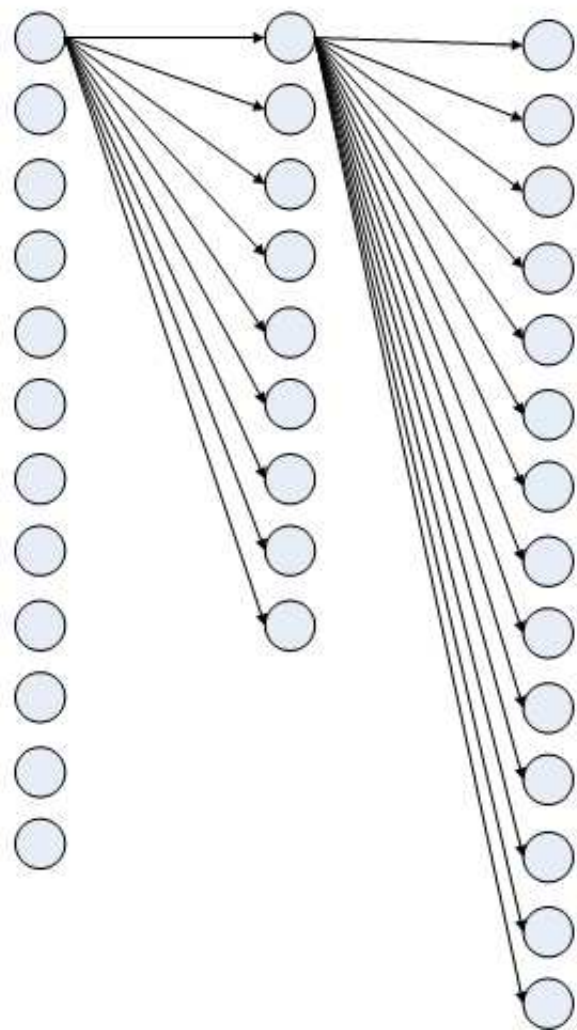


气象趋势综合描述

第一天

第二天

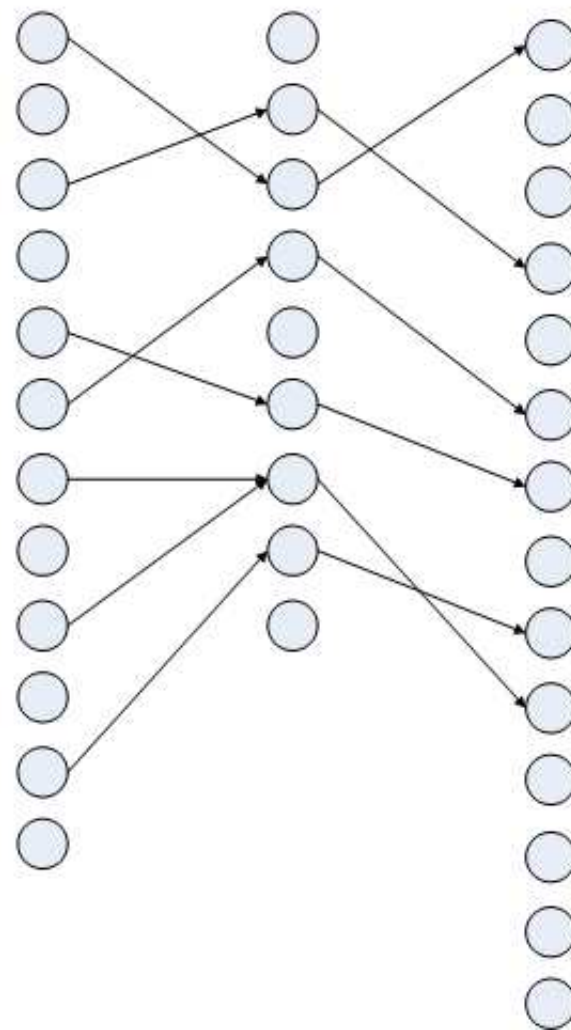
第三天



第一天

第二天

第三天



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



综合预报推理描述

➤ 地理区划分类：

1. 雨雪种类

2. 雨雪量级（比如：小雨、中雨、大雨）和风级

➤ 地理区划融合：

按每个一级区划的三天不同描述匹配，逐个输出

➤ 实例：

未来三天，我国**自西向东**将有一次小到中雨（雪）天气过程。西北东南部 and 东北部、西藏东部和南部、内蒙古西部和东部、东北北部、华北、黄淮大部、江淮、江南、华南、西南等地有**小到中雨**，其中，西北东南部、西藏东部、河南西部、江淮大部、江汉、江南大部、华南大部、云南、贵州大部、四川、重庆等地有**大到暴雨**，其中，江苏南部、安徽中南部、江淮西部、湖北大部、华南西部等地有**大暴雨**。



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY



外置配置文件

- 公报模板
- GIS数据XML接口
- 地理区划描述顺序
- 地理区划一二级包含关系
- 地理区划一三级包含关系
- 气象量级描述规则
- 地理区划描述范围阈值





气象公报自动生成系统DEMO界面

CGEN气象公报自动生成系统

天气公报 | 海洋公报

F:\Test\snow\yr060620.024
F:\Test\snow\yr060720.048
F:\Test\snow\yr060820.072
F:\Test\snow\yr 2010060708024.jpg
F:\Test\snow\yr 2010060708048.jpg
F:\Test\snow\yr 2010060708072.jpg

添加数据

删除

加载数据

阈值调节

请输入0-999之间的整数，阈值转化为千分比形式

局部阈值300

大部阈值700

全部阈值900

省略阈值50

基本信息

预报员:电话:传真:

时间:

2010

年

12

月

25

日

14

时

输出结果路径:

...

生成公报

生成word文件

气象公报生成结果



Thank You!



北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY