# Generating Spatio-Temporal Descriptions in Pollen Forecasts

**Ross Turner, Somayajulu Sripada and Ehud Reiter**
Dept of Computing Science,
University of Aberdeen, UK
{rturner,ssripada,ereiter}@csd.abdn.ac.uk

**Ian P Davy**
Aerospace and Marine International,
Banchory, Aberdeenshire, UK
idavy@weather3000.com

## Abstract

We describe our initial investigations into generating textual summaries of spatio-temporal data with the help of a prototype Natural Language Generation (NLG) system that produces pollen forecasts for Scotland.

## 1 Introduction

New monitoring devices such as remote sensing systems are generating vast amounts of spatio-temporal data. These devices, coupled with the wider accessibility of the data, have spurred large amounts of research into how it can best be analysed. There has been less research however, into how the results of the data analysis can be effectively communicated. As part of a wider research project aiming to produce textual reports of complex spatio-temporal data, we have developed a prototype NLG system which produces textual pollen forecasts for the general public.

Pollen forecast texts describe predicted pollen concentration values for different regions of a country. Their production involves two subtasks; predicting pollen concentration values for different regions of a country, and describing these numerical values textually.In our work, we focus on the later subtask, textual description of spatio-temporally distributed pollen concentration values. The subtask of predicting pollen concentrations is carried out by our industrial collaborator, Aerospace and Marine International (UK) Ltd (AMI).

A fairly substantial amount of work already exists on weather forecast generation. A number of systems have been developed and are currently in commercial use with two of the most notable being FOG (Goldberg et al., 1994) and MultiMeteo (Coch, 1998).

## 2 Knowledge Acquisition

Our knowledge acquisition activities consisted of corpus studies and discussions with experts. We have collected a parallel corpus (69 data-text pairs) of pollen concentration data and their corresponding human written pollen reports which our industrial collaborator has provided for a local commercial television station. The forecasts were written by two expert meteorologists, one of whom provided insight into how the forecasts were written. An example of a pollen forecast text is shown in Figure 1, its corresponding data is shown in table 1. A pollen forecast in the map form is shown in Figure 2.

> *'Monday looks set to bring another day of relatively high pollen counts, with values up to a very high eight in the Central Belt. Further North, levels will be a little better at a moderate to high five to six. However, even at these lower levels it will probably be uncomfortable for Hay fever sufferers.'*

Figure 1: Human written pollen forecast text for the pollen data shown in table 1

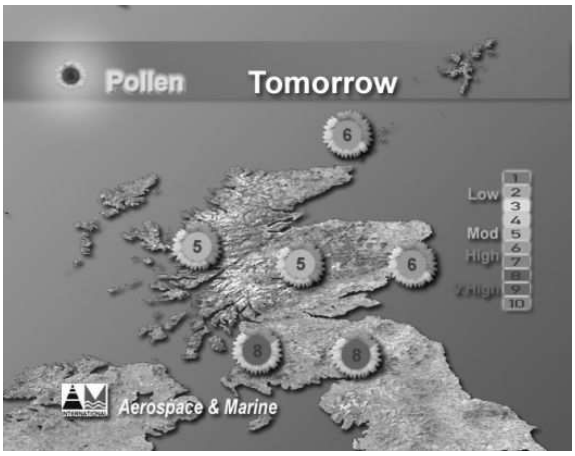


Figure 2: Pollen forecast map for the pollen data shown in table 1

Analysis of a parallel corpus (texts and their underlying data) can be performed in two stages:

- In the first stage, traditional corpus analysis procedure outlined in (Reiter and Dale, 2000) and (Geldof, 2003) can be used to analyse the pollen forecast texts (the textual component of the parallel corpus). This stage will identify the different message types and uncover the sub language of the pollen forecasts.
- In the second stage the more recent analysis methods developed in the SumTime project (Reiter et

| ValidDate | AreaID | Value |
|---|---|---|
| 27/06/2005 | 1 (North) | 6 |
| 27/06/2005 | 2 (North West) | 5 |
| 27/06/2005 | 3 (Central) | 5 |
| 27/06/2005 | 4 (North East) | 6 |
| 27/06/2005 | 5 (South West) | 8 |
| 27/06/2005 | 6 (South East) | 8 |

Table 1: Pollen Concentration Data for Scotland - Input data for Figures 1 and 2

al., 2003) which exploit the availability of the underlying pollen data corresponding to the forecast texts can be used to map messages to input data and also map parts of the sub language such as words to the input data. Due to the fact that we are modeling the task of automatically producing pollen forecast texts from predicted pollen concentration values, knowledge of how to map input data to messages and words/phrases is absolutely necessary. Studies connecting language to data are useful for understanding the semantics of language in a more novel way than the traditional logic-based formalisms (Roy and Reiter, 2005).

We have performed the first stage of the corpus analysis and part of the second stage so far. In the first stage, we abstracted out the different message types from the forecast texts (Reiter and Dale, 2000). These are shown in Table 2. The main two message types are forecast messages and trend messages. The former communicate the actual pollen forecast data (the communicative goal) and the latter describe patterns in pollen levels over time as shown in Figure 3

*'Grass pollen counts continue to ease from the recent high levels'*

Figure 3: A trend message describing a fall in pollen levels

Table 2 also shows three other identified message types. We have ignored both the forecast explanation and general message types in our system development because they cannot be generated from pollen data alone. For example, the explanation type messages explain the weather conditions responsible for the pollen predictions. Hayfever messages in our system are represented as canned text. Examples of a forecast explanation message and hayfever message are shown in Figure 4 and Figure 5 respectively.

From our corpus analysis we have also been able to learn the text structure for pollen forecasts. The forecasts normally start with a trend message and then include a number of forecast messages. Where hayfever messages are present, they normally occur at the end of the forecast.

Due to the fact that the input to our pollen text generator is the pollen data in numerical form, as part of the second stage of the corpus analysis we need to map the input data to the messages. In earlier 'numbers to text' NLG systems such as SumTime (Sripada et al., 2003) and TREND (Boyd, 1998), well known data analysis techniques such as segmentation and wavelet analysis were employed for this task. Since pollen data is spatio-temporal we need to employ spatio-temporal data analysis techniques to achieve this mapping. We describe our method in the next section.

*'Windier and wetter weather over last 24 hours has dampened down the grass pollen count'*

Figure 4: An example forecast explanation message

*'Even though values are mostly low, those sensitive to pollen may still be affected'*

Figure 5: An example hayfever message

Our corpus analysis revealed that forecast texts contain a rich variety of spatial descriptions for a location. For example, the same region could be referred to by it's proper name e.g. *'Suthlerland and Caithness'* or by its' relation to a well known geographical landmark e.g. *'North of the Great Glen'* or simply by its' geographical location on the map e.g. *'the far North and Northwest'*. In the context of pollen forecasts which describe spatio-temporal data, studying the semantics of phrases or words used for describing locations or regions is a challenge. We are currently analysing the forecast texts along with the underlying data to understand how spatial descriptions map to the underlying data using the methods applied in the SumTime project (Sripada et al., 2003).

As part of this analysis, in a seperate study, we asked twenty four further education students in the Glasgow area of Scotland a Geography question. The question asked how many out of four major place names in Scotland did they consider to be in the south west of the country. The answers we got back were very mixed with a sizeable number of respondents deciding that the only place we considered definitely not to be in the south west of Scotland was in fact there.

## 3 Spatio-temporal Data Analysis

We have followed the pipeline architecture for text generation outlined in (Reiter and Dale, 2000). The microplanning and surface realisation modules from the Sumtime project (Sripada et al., 2003) have largely been reused. We have developed new data analysis and document planning modules for the system and describe the data analysis module in the rest of this section. The data analysis module performs segmentation and trend detection on the data before providing the results as input to the Natural Language Generation Sys-

| Message Type | Data Dependency | Corpus Coverage |
|---|---|---|
| Forecast | Pollen data for day of forecast | 100% |
| Trend | Past/Future pollen forecasts | 54% |
| Forecast Explanation | Weather forecast for day of forecast | 35% |
| Hayfever | Pollen levels affect hay fever | 23% |
| General | General Domain Knowledge | 17% |

Table 2: Message Categorisation of the Pollen Corpus

tem. An example of the input data to our system is shown in Table 1. Our data analysis is based on three steps:-

1. segmentation of the geographic regions by their non-spatial attributes (pollen values)
2. further segmentation of the segmented geographic regions by their spatial attributes (geographic proximity)
3. detection of trends in the generalised pollen level for the whole region over time

### 3.1 Segmentation

The task of segmentation consists of two major subtasks, clustering and classification (Miller and Han, 2001). Spatial clustering involves grouping objects into similar subclasses, whereas spatial classification involves finding a description for those subclasses which differentiates the clustered objects from each other (Ester et al., 1998).

Pollen values are measured on a scale of 1 to 10(low to very high). We defined 4 initial categories for segmentation, these are:-

1. VeryHigh - {8,9,10}
2. High - {6,7}
3. Moderate - {4,5}
4. Low - {1,2,3}

These categories proved rather rigid for our purposes. This was due to the fact that human forecasters take a flexible approach to classifying pollen values. For example, in the corpus the pollen value of 4 could be referred to as both a moderate level of pollen and a low-to-moderate level of pollen. This lead us to define 3 further categories which are derived from our 4 initial categories:-

5. LowModerate - {3,4}
6. ModerateHigh - {5,6}
7. HighVeryhigh - {7,8}

Thus, the initial segmentation of data carried out by our system is a two stage process. Firstly regions are clustered into the initial four categories by pollen value. The second stage involves merging adjacent categories that only contain regions with adjacent values. For example if we take the input data from Table 1, after the first stage we have the sets:-

- {{AreaID=2,Value=5},{AreaID=3,Value=5}}
- {{AreaID=1,Value=6},{AreaID=4,Value=6}}
- {{AreaID=5,Value=8},{AreaID=6,Value=8}}

In stage two we create the union of the moderate and high sets to give:-

- {{AreaID=1,Value=6},{AreaID=2,Value=5}, {AreaID=3,Value=5},{AreaID=4,Value=6}}
- {{AreaID=5,Value=8},{AreaID=6,Value=8}}

Although this initial segmentation could be accomplished all in one step, completing it in two steps provided a more simple software engineering solution.

We can now carry out further segmentation of these sets according to their spatial attributes. In our set of regions with ModerateHigh pollen levels we can see that AreaIDs 1,2,3,4 are in fact all spatial neighbours. The north, north east and north west regions can be described spatially as the northern part of the country. Therefore we can now say that *'Pollen levels are at a moderate to high 5 or 6 in the northern and central parts of the country'* . Similarly, as the two members of our set containing regions with VeryHigh pollen levels are also spatial neighbours we can also say that *'Pollen levels are at a very high level 8 in the south of the country'*. This process now yields the following two sets:-

- {{AreaID=1234,Value=[5,6]}}
- {{AreaID=56,Value=[8]}}

Our two sets we have now created can now be passed to the Document Planner were they will be encapsulated as individual Forecast messages.

### 3.2 Trend Detection

Trend detection in our system works by generalising over all sets created by segmentation. From our two sets we can say that generally pollen levels are high over the whole of Scotland. Looking at the previous days forecast we can detect a trend by comparing the two generalisations. If the previous days forecast was also high we can say *'pollen levels remain at the high*

*levels of yesterday'*. By looking further back, and if those previous days were also high, we can say *'pollen levels remain at the high levels of recent days'*. If the previous days forecast was low, we can say *'pollen levels have increased from yesterdays low levels'*. Our data analysis module then conveys the information that there is a relation between the general pollen level of today and the general pollen level of some recent timescale to the Document Planner, which then encapsulates the information as a Trend message.

After the results of data analysis have been input into the NLG pipeline the output in Figure 6 is produced.

> *'Grass pollen levels for Monday remain at the moderate to high levels of recent days with values of around 5 to 6 across most parts of the country. However, in southern areas, pollen levels will be very high with values of 8.'*

Figure 6: The output text from our system for the input data in Table 1

## 4 Evaluation

A demo of the pollen forecasting system can be found on the internet at [1]. The evaluation of the system is being carried out in two stages. The first stage has used this demo to obtain feedback from expert meteorologists at AMI. We found the feedback on the system to be very positive and hope to deploy the system for the next pollen season. Two main areas identified for improvement of the generated texts:-

- Use of a more varied amount of referring expressions for geographic locations.
- An ability to vary the length of the text dependent on the context it was being used, i.e in a newspaper or being read aloud.

These issues will be dealt with subsequent releases of the software. The second and more thorough evaluation will be carried out when the system is deployed.

## 5 Further Research

The current work on pollen forecasts is carried out as part of **RoadSafe**[2] a collaborative research project between University of Aberdeen and Aerospace and Marine International (UK) Ltd. The main objective of the project is to automatically generate road maintenance instructions to ensure efficient and correct application of salt and grit to the roads during the winter. The core requirement of this project is to describe spatio-temporal data of detailed weather and road surface temperature predictions textually. In a previous research project SumTime (Sripada et al., 2003) we have developed techniques for producing textual summaries of time series data. In **RoadSafe** we plan to extend these techniques to generate textual descriptions of spatio-temporal data. Because the spatio-temporal weather prediction data used in road maintenance applications is normally of the order of a megabyte, we initially studied pollen forecasts which are based on smaller spatio-temporal data sets. We will apply the various techniques we have learnt from the study of pollen forecasts to the spatio-temporal data from the road maintenance application.

## 6 Summary

Automatically generating spatio-temporal descriptions involves two main subtasks. The first subtask focuses on the spatio-temporal analysis of the input data to extract information required by the different message types identified in the corpus analysis. The second subtask is to find appropriate linguistic form for the spatial location or region information.

[1] www.csd.abdn.ac.uk/∼rturner/cgi_bin/pollen.html

[2] www.csd.abdn.ac.uk/∼rturner/RoadSafe/

## References


S. Boyd. 1998. Trend: a system for generating intelligent descriptions of time-series data. In *IEEE International Conference on Intelligent Processing Systems (ICIPS1998)*.

J. Coch. 1998. Multimeteo: multilingual production of weather forecasts. *ELRA Newsletter*, 3(2).

M. Ester, A. Frommelt, H. Kriegel, and J. Sander. 1998. Algorithms for characterization and trend detection in spatial databases. In *KDD*, pages 44–50.

S. Geldof. 2003. Corpus analysis for nlg. citeseer.ist.psu.edu/583403.html.

E. Goldberg, N. Driedger, and R. Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

H. J. Miller and J. Han. 2001. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

E. Reiter, S. Sripada, and R. Robertson. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research*, 18:491–516.

D. Roy and E. Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167:1–12.

S. Sripada, E. Reiter, and I. Davy. 2003. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6:4–10.