A Case Study: NLG meeting Weather Industry Demand for Quality and Quantity of Textual Weather Forecasts

Somayajulu G Sripada, Neil Burnett and Ross Turner
Arria NLG Plc
{yaji.sripada,neil.burnett,
ross.turner}@arria.com

John Mastin and Dave Evans
Met Office
{john.mastin,
dave.evans}@metoffice.gov.uk

Abstract

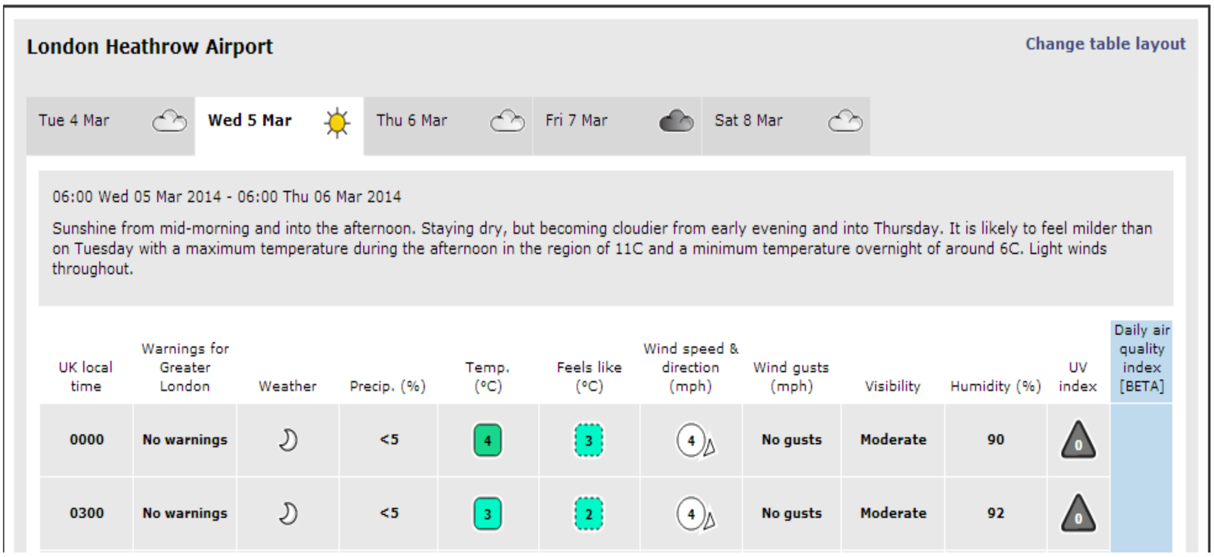
In the highly competitive weather industry, demand for timely, accurate and personalized weather reports is always on the rise. In this paper we present a case study where Arria NLG and the UK national weather agency, the Met Office came together to test the hypothesis that NLG can meet the quality and quantity demands of a real-world use case.

1 Introduction

Modern weather reports present weather prediction information using tables, graphs, maps, icons and text. Among these different modalities only text is currently manually produced, consuming significant human resources. Therefore releasing meteorologists' time to add value elsewhere in the production chain without sacrificing quality and consistency in weather reports is an important industry goal. In addition, in order to remain competitive, modern weather services need to provide weather reports for any geolocation the end-user demands. As the quantity of required texts increases, manual production becomes humanly impossible. In this paper we describe a case study where data-to-text NLG techniques have been applied to a real-world use case involving the UK national weather service, the Met Office. In the UK, the Met Office provides daily weather reports for nearly 5000 locations which are available through its public website. These reports contain a textual component that is not focused on the geo-location selected by the end-user, but instead describes the weather conditions over a broader geographic region. This is done partly because the time taken to manually produce thousands of texts required would be in the order of weeks rather than minutes. In this case study a data-to-text NLG system was built to demonstrate that the site-specific data could be enhanced with site-specific text for nearly 5000 locations. This system, running on a standard desktop, was tested to produce nearly 15000 texts (forecasts for 5000 locations for 3 days into the future) in less than a minute. After internally assessing the quality of machine-generated texts for nearly two years, the Met Office launched the system on their beta site (http://www.metoffice.gov.uk/public/weather/forecast-data2text/) in December 2013 for external assessment. A screenshot of the forecast for London Heathrow on 5th March 2014 is shown in Figure 1. In this figure, the machine-generated text is at the top of the table. Ongoing work has extended the processing capabilities of this system to handle double the number of locations and an additional two forecast days. It has been found that the processing time scales linearly.

2 Related Work

Automatically producing textual weather forecasts has been the second favorite application for NLG, with 15 entries on Bateman and Zock's list of NLG application domains (the domain of medicine comes on top with 19 entries) [Bateman and Zock, 2012]. NLG applications in the weather domain have a long history. FOG was an early landmark NLG system in the domain of weather reports [Goldberg et al, 1994]. Working as a module of the Forecast Production Assistant (FPA), FOG was operationally deployed at Environment Canada to produce weather reports for the general public and also for marine users in both English and French. Using sampling and smoothing over space and time, FOG reduces raw data into a few significant events which are then organized and realized in textual form. MULTIMETEO is another industry deployed multi-lingual weather report generator [Coch 1998]. The focus of MULTIMETEO is 'interactive generation via knowledge administration'.



London Heathrow Airport

06:00 Wed 05 Mar 2014 - 06:00 Thu 06 Mar 2014

Sunshine from mid-morning and into the afternoon. Staying dry, but becoming cloudier from early evening and into Thursday. It is likely to feel milder than on Tuesday with a maximum temperature during the afternoon in the region of 11C and a minimum temperature overnight of around 6C. Light winds throughout.

UK local time	Warnings for Greater London	Weather	Precip. (%)	Temp. (°C)	Feels like (°C)	Wind speed & direction (mph)	Wind gusts (mph)	Visibility	Humidity (%)	UV index	Daily air quality index [BETA]
0000	No warnings	☽	<5	4	3	4	No gusts	Moderate	90	0	
0300	No warnings	☽	<5	3	2	4	No gusts	Moderate	92	0	

Figure 1. Screenshot of Text-Enhanced Five-day Weather Forecast for London Heathrow on 5 March 2014 showing only part of the data table

Expert forecasters post-edit texts (interactivity) in their native language and this knowledge is then reused (knowledge administration) for automatically generating texts in other languages. It is claimed that such interactive generation is better than using machine translation for multilingual outputs. SUMTIME-MOUSAM is yet another significant weather report generator that was operationally deployed to generate forecasts in English for oil company staff supporting oil rig operations in the North Sea [Sripada et al, 2003a]. Adapting techniques used for time series segmentation, this project developed a framework for data summarization in the context of NLG [Sripada et al, 2003b]. This time series summarization framework was later extended to summarizing spatio-temporal data in the ROADSAFE system [Turner et al, 2008]. ROADSAFE too was used in an industrial context to produce weather reports (including text in English and a table) for road maintenance in winter months. The NLG system reported in the current case study builds upon techniques employed by earlier systems, particularly SUMTIME-MOUSAM and ROADSAFE.

The main dimension on which the application described in this paper differs most from the work cited previously is the quantity of textual weather forecasts that are generated. Previous work has either focused on summarising forecast sites collectively (in the case of FOG and ROADSAFE), been limited in the number of sites forecast for (15 in the case of MULTIMETEO) or limited in geographic extent (SUMTIME-MOUSAM concentrated on oil rig operations in the North Sea). This aspect of the system, amongst others, posed a number of challenges discussed in Section 3.

3 System Description

For reasons of commercial sensitivity, the system description in this section is presented at an abstract level. At the architecture level, our system uses the Arria NLG Engine that follows the standard five stage data-to-text pipeline [Reiter, 2007]. The system integrates application specific modules with the generic reusable modules from the underlying engine. Input to the system is made up of three components:

1. Weather prediction data consisting of several weather parameters such as temperature, wind speed and direction, precipitation and visibility at three hourly intervals;
2. Daily summary weather prediction data consisting of average daily and nightly values for several weather parameters as above; and
3. Seasonal averages (lows, highs and mean) for temperature.

Because the system is built on top of the Arria NLG Engine, input data is configurable and not tied to file formats. The system can be configured to work with new data files with equivalent weather parameters as well as different forecast periods. In other words, the system is portable in principle for other use cases where site-specific forecasts are required from similar input data.

3.1 Expressing Falling Prediction Quality for Subsequent Forecast Days

As stated above, the system can be configured to generate forecast texts for a number of days into the future. Because prediction accuracy reduces going into the future, the forecast text on day 1 should be worded differently from subsequent days where the prediction is relatively more uncertain. An example output text for day 1 is shown in Figure 2 while Figure 3 shows the day 3 forecast. Note the use of ‘expected’ to denote the uncertainty around the timing of the temperature peak.

Staying dry and predominantly clear with only a few cloudy intervals through the night. A mild night with temperatures of 6C. Light winds throughout.

Figure 2. Example output text for day 1

Cloudy through the day. Mainly clear into the night. Highest temperatures expected during the afternoon in the region of 12C with a low of around 6C during the night. Light to moderate winds throughout.

Figure 3. Example output text for day 3

3.2 Lack of Corpus for System Development

A significant feature of the system development has been to work towards a target text specification provided by experts rather than extract such a specification from corpus texts, as is generally the case with most NLG system development projects. This is because expert forecasters do not write the target texts regularly; therefore, there is no naturally occurring target corpus. However, because of the specialized nature of the weather sublanguage (Weatherese), which has been well studied in the NLG community [Goldberg et al, 1994, Reiter et al 2005, Reiter and Williams 2010], it was possible to supplement text specifications obtained from experts. In addition, extensive quality assessment (details in section 3.4) helped us to refine the system output to the desired levels of quality.

3.3 Achieving Output Quantity

The main requirements of the case study have been 1) build a NLG capability that produces the quantity of texts required and 2) achieve this quantity without sacrificing the quality expected from the Met Office. As stated previously, the quantity requirement has been met by generating 15,000 texts in less than a minute, without need for high end computing infrastructure or parallelization. Figure 4 is a box plot showing character lengths of forecast texts for an arbitrary set of inputs. The median length is 177 characters. The outliers, with length 1.5 times the interquartile range (1.5 * 134 = 201 characters) above the upper quartile or below the lower quartile, relate to sites experiencing particularly varied weather conditions. Feedback on the appropriateness of the text lengths is discussed in Section 3.4.

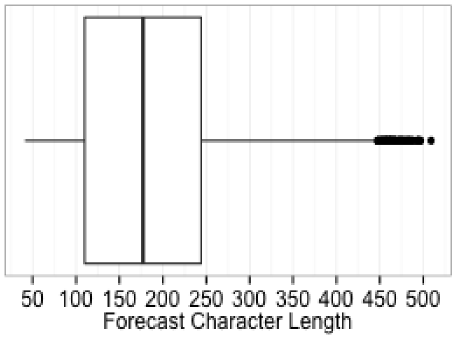


Figure 4. Boxplot of forecast character length

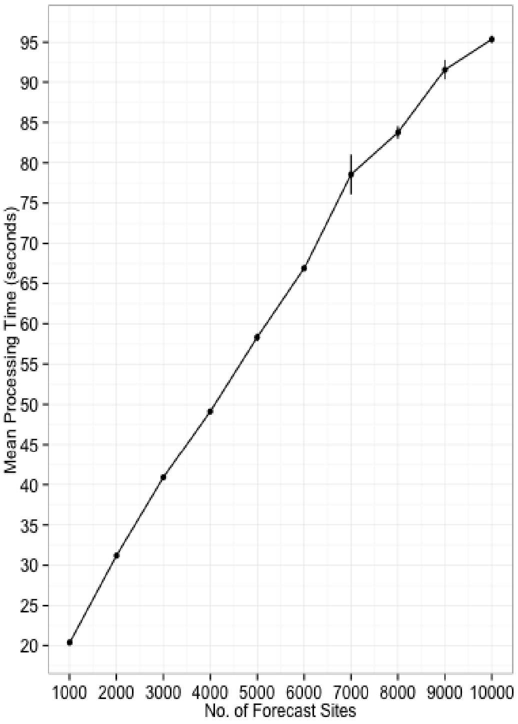


Figure 5. System Processing Time

The system has recently been extended to generate 50,000 texts without loss of performance. This extension has doubled the number of sites processed to 10,000 and extended the forecast to 5 days. It has also increased the geographic extent of the system from UK only to worldwide, discussed in Section 3.5. The plot in Figure 5 shows the relationship between processing time and the addition of new forecast sites. The results were obtained over 10 trials using a MacBook Pro 2.5 GHz Intel Core i5, running OS X 10.8 with 4GB of RAM.

3.4 Achieving Output Quality

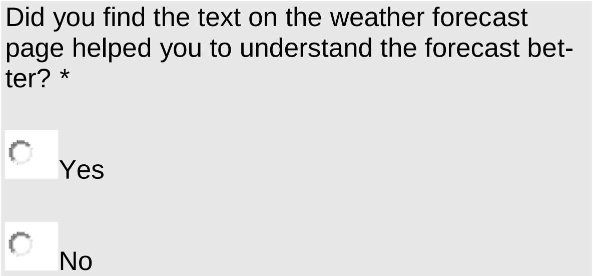
Achieving the required text quality was driven by expert assessment of output texts that occurred over a period of two years. This is because experts had to ensure that the system output was assessed over the entire range of weather conditions related to seasonal variations over the course of a year. The following comment about the output quality made by a Met Office expert summarizes the internal feedback:

"They were very, very good and I got lots of verbal feedback to that affect from the audience afterwards. Looking back after the weekend, the forecasts proved to be correct too! I've been looking at them at other times and I think they're brilliant."

After successfully assessing the output quality internally, the Met Office launched the system on the Invent part of their website to collect end-user assessments. Invent is used by the Met Office to test new technology before introducing the technology into their workflows. With the help of a short questionnaire[1] that collects assessment of those end-users that use weather information for decision-making, quality assessment is ongoing. The questionnaire had three questions related to quality assessment shown in Figures 6-8. In the rest of the section we describe the results of this questionnaire based on 35 responses received between 1st January 2014 and 6th March 2014.

The first question shown in Figure 6 relates to assessing the usefulness of textual content in helping the end-user understand a weather report better. Out of the 35 respondents, 34 (97%) answered '*yes*' and 1 (3%) answered '*no*' for the question in Figure 6. The second question shown in Figure 7 relates to assessing if the text size is optimal for this use case. Here, out of the 35 respondents, 26 (74%) felt the text is *'about right'* size, 7 (20%) felt it is either *'too short'* or *'too long'* and 2 (6%) were *'unsure'*. The third question shown in Figure 8 relates to finding out if the end-user might want a forecast that includes textual content. Here, 32 (91%) wanted textual content while 3 (9%) did not want it.

The Met Office is currently evaluating the new capability based upon the feedback received and how it can be applied to meet the demands of users across their portfolio of products.

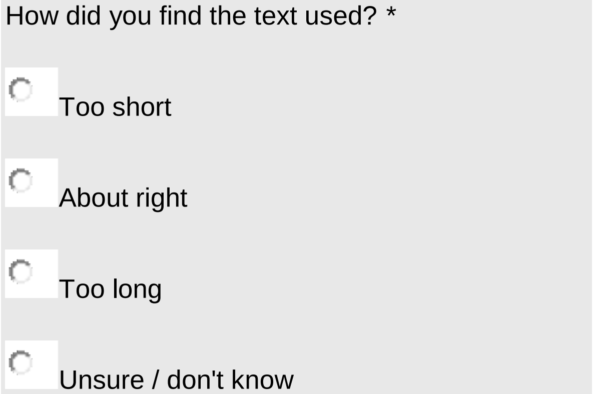


Did you find the text on the weather forecast page helped you to understand the forecast better? *

○ Yes

○ No

Figure 6. Question about textual content helping the end-user understand the forecast better



How did you find the text used? *

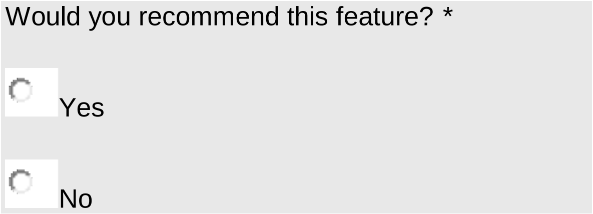
○ Too short

○ About right

○ Too long

○ Unsure / don't know

Figure 7. Question about length of the forecast text



Would you recommend this feature? *

○ Yes

○ No

Figure 8. Question about the end-user's opinion on textual content as part of a weather report

The questionnaire also asked for free text comments. An example of one such comment is:

"Succinct and clear text. Contains all the important features and is well presented. Saves us having to summarise the visual descriptions ourselves (or rather helps to shape our conclusions about the 24 hour weather pattern)."

A big challenge during the development of such a system is providing quality assurance

[1] http://www.metoffice.gov.uk/invent/feedback

when generating such a large volume of texts. A number of automated checks had to be applied to the complete output during system testing as well as targeted sampling of input data to produce a representative sample of outputs for manual assessment.

3.5 Extending the Geographic Extent

Extending the scope of the system from UK-only sites to handling worldwide locations brings subtle challenges in addition to scaling the system, principally:

1. handling time zone changes; and
2. adapting to different climates.

In the case of point 1 above, time descriptions can become ambiguous where the sunrise and sunset time vary across geographies. Such times need to be carefully observed to avoid generating words such as “sunny” after dark. For point 2, general terminologies relating to description of temperatures cannot be universally applied across locations. For example, the meaning of terms such as “cool” differs at locations within the tropics versus locations north (or south) of 45 degrees of latitude.

4 Conclusion

We have presented a case study describing an application of NLG technology deployed at the Met Office. The system has been developed to meet the text production requirements for thousands of forecast locations that could not have been sustainable with human resources. The software can write a detailed five-day weather forecast for 10,000 locations worldwide in under two minutes. It would take a weather forecaster months to create the equivalent quantity of output.

In meeting the requirements of this particular use case a number of challenges have had to be met. Principally, these challenges have been focused upon processing speed and output text quality. While we have managed to achieve the required processing performance relatively quickly without the need for large amounts of computing resources or high-end computing infrastructure, ensuring the necessary output quality has been a longer process due to the high operating standards required and the high resource cost of quality assurance when delivering texts at such scale.

This application of NLG technology to site-specific weather forecasting has potential for a number of enhancements to the type of weather services that may be provided in the future, most notably the opportunity for very geographically localized textual forecasts that can be updated immediately as the underlying numerical weather prediction data is produced.

References

E. Goldberg, N. Driedger, and R. Kittredge. Using Natural-Language Processing to Produce Weather Forecasts. IEEE Expert, 9(2):45--53, 1994.

J. Coch. Interactive generation and knowledge administration in MultiMeteo. In Proceedings of the Ninth International Workshop on Natural Language Generation, pages 300--303, Niagara-on-the-lake, Ontario, Canada, 1998. software demonstration.

Bateman J and Zock M, (2012) Bateman/Zock list of NLG systems, http://www.nlg-wiki.org/systems/.

S. Sripada, E. Reiter, and I. Davy, (2003a) ‘SumTime-Mousam: Configurable Marine Weather Forecast Generator’, Expert Update, 6(3), pp 4-10, (2003)

S. Sripada, E. Reiter, J. Hunter and J. Yu (2003b). Generating English Summaries of Time Series Data using the Gricean Maxims. In Proceedings of KDD 2003, pp 187-196.

E. Reiter, S. Sripada, J. Hunter, J. Yu and Ian Davy (2005). Choosing Words in Computer-Generated Weather Forecasts. Artificial Intelligence. 167(1-2):137-169

E. Reiter (2007). An architecture for data-to-text systems, In ENLG 07, pp97-104.

Reiter, Ehud and Williams, Sandra (2010). Generating texts in different styles. In: Argamon, Shlomo; Burns, Kevin and Dubnov, Shlomo eds. The Structure of Style: Algorithmic Approaches to Manner and Meaning. Heidelberg: Springer, pp. 59–78

E. Reiter, S. Sripada, J. Hunter, J. Yu, and Ian Davy(2005). Choosing words in computer-generated weather forecasts. Artificial Intelligence. 167(1-2):137-169 (2005)

R. Turner, S. Sripada, E. Reiter, & I. Davy (2008). Using spatial reference frames to generate grounded textual summaries of geo-referenced data. Proceedings of the INLG 2008, Salt Fork, Ohio.