2.2

COMPUTATIONAL LINGUISTICS AND THE COMMUNICATION OF WEATHER FORECASTS

Harvey Stern*
School of Earth Sciences,
University of Melbourne, Victoria, Australia

1. INTRODUCTION

The *Association for Computational Linguistics* (ACL) defines the term *Computational Linguistics* as the scientific study of language from a computational perspective.

The ACL notes that computational linguists are interested in providing computational models of various kinds of linguistic phenomena and that these models may be:

- Knowledge-based (hand-crafted); or,
- Data-driven (statistical or empirical).

2. BACKGROUND

The ACL further notes that work in computational linguistics is, in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon, whilst, in other cases, the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system.

3. PURPOSE

The purpose of the current paper is to presents an analysis of the words used in a 12-year data set (2005-2017) of précis weather forecasts for Melbourne, Australia.

4. METHODOLOGY

The analysis strategy has three components:

Firstly, to study the overall frequency of occurrence, of particular words and phrases;

Secondly, to note any significant trends over the period, in the nature of the language utilised to communicate the weather forecast information.

Thirdly, to establish how one might best combine textual components of weather forecasts with numerical components (for example, precipitation amount and probability).

* *Corresponding author address: Harvey Stern, School of Earth Sciences, University of Melbourne, Parkville, Victoria, Australia, 3010;* e-mail: hstern@unimelb.edu.au

What is being sought here is an optimal way of combining textual components of weather forecasts with numerical components (for example, precipitation amount and probability) that leads to an overall enhancement in the accuracy of the predictions.

5. RESULTS

5.1 *Frequency of occurrence*

The ten most frequently occurring Day-1 précis weather forecasts issued by the Australian Bureau of Meteorology for Melbourne over the twelve years were:

PARTLY CLOUDY (8.4%),
SHOWER OR TWO (7.8%),
MOSTLY SUNNY (6.6%),
FINE (6.2%),
SUNNY (5.3%),
FEW SHOWERS (3.3%),
A FEW SHOWERS (3.1%),
SHOWER OR TWO CLEARING (2.5%),
BECOMING FINE (2.0%),
POSSIBLE SHOWER (1.7%).

Figure 1 expands the foregoing list by depicting the frequency distribution of the 50 most common précis weather forecasts.

5.2 *Trends in language use*

The most dramatic change in the language utilised relates to FINE which was used on 20% of occasions during the first year, but was completely absent during the last year.

By contrast, the précis PARTLY CLOUDY, which was not used at all during the first year, was used on 16% of occasions during the final year.

5.3 *Combining textual and numerical components*

The textual components of weather forecasts (e.g. A FEW SHOWERS) are blended with the official numerical components (e.g. Probability (likelihood) of Precipitation occurrence: e.g. *70% (PoP)*, and amount: e.g. *2 to 5 mm: Low Amount to High Amount*). This is done in order to establish whether such blending has the potential to enhance accuracy.

5.3.1 *Probability of Precipitation (PoP)*

A multiple linear regression relationship is derived between the words utilised and the subsequent occurrence (or non-occurrence) of precipitation. The equation so derived is found to explain 42.7% of the variance of the observed subsequent precipitation likelihood (PoP).

Figure 2.1 illustrates the associated partial regression coefficients, t statistics and P-values (their significance).

Figure 2.1 shows that the words in précis weather forecasts most highly correlated with subsequent precipitation likelihood are:

RAIN
SHOWERS
SHOWER
DRIZZLE
THUNDER

By contrast, Figure 2.1 shows that the words in précis weather forecasts least correlated with subsequent precipitation likelihood are:

LITTLE
CHANCE
FEW
CLEARING
LATE

In contrast to the equation so derived being found to explain 42.7% of the variance of the observed subsequent precipitation likelihood (PoP), the official PoP is found to explain 45.6%, the official LOW amount is found to explain 47.2%, and the official HIGH amount is found to explain 53.4%.

A multiple linear regression relationship is now derived between the aforementioned numerical estimates and the subsequent occurrence (or non-occurrence) of precipitation. Figure 2.2 illustrates the associated partial regression coefficients, t statistics and P-values (their significance).of the equation so derived. The equation explains 54.3% of the variance of the observed subsequent precipitation likelihood, an increase, albeit a slight increase, on that explained by any of its components.

5.3.2 *Amount of Precipitation*

A multiple linear regression relationship is derived between the words utilised and the subsequent amount of precipitation.

Figure 2.3 illustrates the associated partial regression coefficients, t statistics and P-values (their significance).

Figure 2.3 shows that the words in précis weather forecasts most highly correlated with precipitation amount are:

RAIN
SHOWERS
SHOWER
HEAVY
THUNDER

By contrast, Figure 2.3 shows that the words in précis weather forecasts least correlated with precipitation likelihood are:

LITTLE
FEW
CHANCE
CLEARING
LATE

In contrast to the equation so derived being found to explain 53.0% of the variance of the observed subsequent precipitation amount, the official PoP is found to explain 47.6%, the official LOW amount is found to explain 59.9%, and the official HIGH amount is found to explain 64.4%.

A multiple linear regression relationship is now derived between the aforementioned numerical estimates and the subsequent amount of precipitation. Figure 2.4 illustrates the associated partial regression coefficients, t statistics and P-values (their significance).of the equation so derived. The equation explains 64.7% of the variance of the observed subsequent precipitation likelihood, an increase, albeit only a very slight increase, on that explained by any of its components.

6. CONCLUDING REMARKS

The frequency distribution of various words used in the official weather forecasts has been established, and it has been shown that their usage varies over time as different words become less or more ‘fashionable’.

The analysis approach described here is shown to enhance the accuracy of the numerical components of the official forecasts, albeit only slightly.

However, the approach readily achieves the identification of how the individual components of the official forecasts, and also the words utilised therein, are related to both the amount and likelihood of subsequent precipitation.

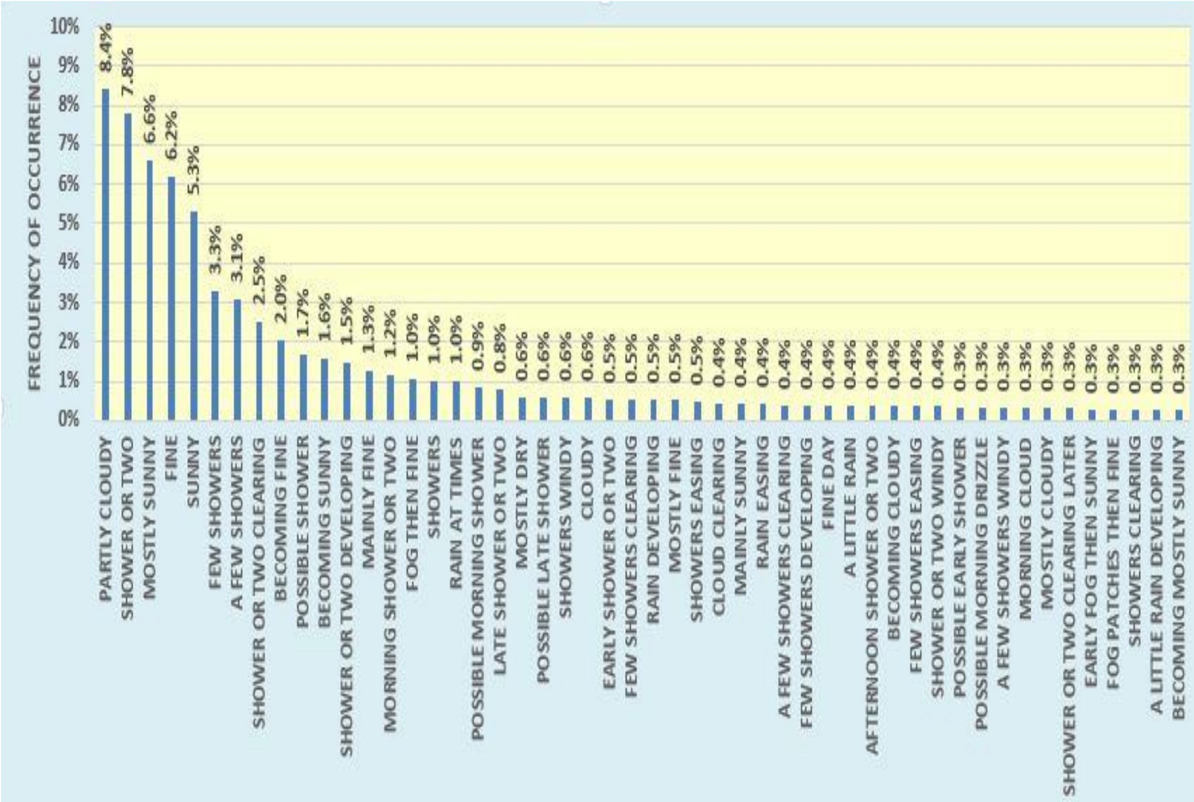

Figure 1 The frequency distribution of the 50 most common précis weather forecasts.

VARIABLE	*Coefficients*	*t Stat*	*P-value*
CONSTANT	9.3914	7.99	8.5E-16
RAIN	73.6175	33.15	1.5E-215
SHOWER	35.7348	22.68	4.7E-108
SHOWERS	33.4854	14.94	1.4E-49
DRIZZLE	26.7889	6.80	6.0E-12
THUNDER	16.5787	5.47	2.4E-08
BECOMING	8.2476	3.39	3.5E-04
EASING	5.7797	2.10	1.8E-02
HEAVY	16.1928	1.31	9.6E-02
CLOUD	-2.1751	-1.18	1.2E-01
FOG	-4.7680	-1.70	4.5E-02
FINE	-3.9270	-2.39	8.4E-03
LATE	-4.3014	-2.52	5.9E-03
CLEARING	-4.8558	-2.66	4.0E-03
FEW	-6.6368	-2.68	3.7E-03
CHANCE	-13.7656	-6.33	1.3E-10
LITTLE	-24.6880	-7.65	1.3E-14

Figure 2.1 The multiple linear regression relationship between the likely subsequent occurrence (probability) of precipitation (PoP) and the words utilised.

VARIABLE	Coefficients	t Stat	P-value
CONSTANT	4.2738	1.11	1.3E-01
SQRT(HIGH)	23.2366	5.76	4.7E-09
PoP	0.7085	5.19	1.2E-07
Predicted*	0.2384	1.77	3.8E-02
SQRT(LOW)	0.3642	0.10	4.6E-01
LOW	-0.3749	-0.22	4.1E-01
SQRT(Pred)	-1.1045	-0.71	2.4E-01
SQRT(PoP)	-2.2314	-2.09	1.9E-02
HIGH	-2.8180	-2.86	2.2E-03

Figure 2.2 The multiple linear regression relationship between the likely subsequent occurrence (probability) of precipitation and the official PoP (and its SQRT), the official LOW amount (and its SQRT), the official HIGH amount (and its SQRT), and the PoP from the equation presented at Figure 2 *(Predicted/Pred)* and its SQRT.

VARIABLE	Coefficients	t Stat	P-value
CONSTANT	0.1417	5.43	3.0E-08
RAIN	2.0797	42.15	0.0E+00
SHOWERS	0.9540	19.16	5.8E-79
SHOWER	0.4994	14.27	1.8E-45
HEAVY	2.5934	9.42	3.4E-21
THUNDER	0.5546	8.24	1.1E-16
DRIZZLE	0.4940	5.64	8.9E-09
EASING	0.3008	4.93	4.3E-07
BECOMING	0.0888	1.64	5.0E-02
FOG	-0.0260	-0.42	3.4E-01
CLOUD	-0.0696	-1.70	4.4E-02
FINE	-0.1144	-3.13	8.7E-04
LATE	-0.1515	-3.99	3.3E-05
CLEARING	-0.1670	-4.11	2.0E-05
CHANCE	-0.2616	-5.42	3.2E-08
FEW	-0.4505	-8.19	1.6E-16
LITTLE	-0.8675	-12.10	1.9E-33

Figure 2.3 The multiple linear regression relationship between the likely subsequent amount of precipitation and the words utilised.

VARIABLE	Coefficients	t Stat	P-value
CONSTANT	0.0630	1.47	7.1E-02
Predicted*	0.4335	6.11	6.0E-10
HIGH	0.0766	3.77	8.5E-05
SQRT(HIGH)	0.2686	3.24	6.0E-04
PoP	0.0090	3.22	6.5E-04
SQRT(LOW)	0.1237	1.65	4.9E-02
SQRT(PoP)	-0.0430	-1.95	2.5E-02
LOW	-0.0892	-2.54	5.6E-03
SQRT(Pred)	-0.3394	-3.12	9.2E-04

Figure 2.4 The multiple linear regression relationship between the likely subsequent amount of precipitation and the official PoP (and its SQRT), the official LOW amount (and its SQRT), the official HIGH amount (and its SQRT), and the amount from the equation presented at Figure 2.3 *(Predicted/Pred)* and its SQRT.