SUMTIME-METEO: Parallel Corpus of Naturally Occurring Forecast Texts and Weather Data (Revised 14 Nov 2005)

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

Abstract

In this report, we describe SUMTIME-METEO, a parallel corpus of naturally occurring weather data and their corresponding forecast texts. The corpus has 1045 parallel data-text units and is available as a Microsoft Access database.

1. Introduction

The goal of SUMTIME project was to develop general techniques for generating text summaries of time-series data. In order to achieve this objective we studied how human experts wrote text summaries of time-series data, and in particular gathered a corpus of human-written weather forecasts (marine forecasts intended for offshore oil rigs in the North Sea), produced by staff at a weather forecasting company. We also gathered the numerical data (that is, numerical predictions of wind speed, temperature, etc) that the human forecasters examined when they wrote the forecasts. The SUMTIME-METEO corpus (often just referred to as the SUMTIME corpus) contains 1045 human written forecasts and associated data files, written between 26-June-2000 and 10-May-2002.

We call SUMTIME-METEO a parallel text-data corpus because it contains both non-linguistic input data (in our case, numerical weather predictions) and output texts written from this data. This is an analogy with parallel text-text (bitext) corpora used in machine translation (Brown et al, 1990)., which contain "input" texts in the source language and corresponding human-written "output" texts in the target language.

For more information on the SumTime project, see our web page (http://www.csd.abdn.ac.uk /research/sumtime), which includes a publication list. Key publications which are related to the use of the SUMTIME-METEO corpus are

- Reiter, Sripada, Robertson (2003), which describes our general knowledge acquisition methodology for building NLG systems, including corpus analysis
- Sripada, Reiter, Davy, Nilssen (2004), which describes the actual weather-forecast generator that we built and how it is used operationally.
- Reiter, Sripada, Hunter, Yu, Davy (2005), which describes how we used the SUMTIME-METEO corpus to build the lexical-choice module of our weather forecast generator, and an experiment which suggests that our computer-generated forecasts may be better in some ways than human-written forecasts.

- Sripada, Reiter, Hawizy (2005), which describes another corpus we have created, which contains computer-generated forecasts which have been post-edited by human forecasters.

The only paper which specifically is about the SUMTIME-METEO corpus is Sripada, Reiter, Hunter, and Yu (2003). We suggest that people cite this paper if they want an official reference for SUMTIME-METEO.

For an example of how other researchers have used SUMTIME-METEO, see Belz (2005).

2. Conditions for Using this Corpus

Please note that the below conditions of use only apply to the version of the corpus released in 2005. More stringent conditions apply to an earlier version of the corpus which we distributed to a few people in earlier years (because this corpus was less anonymised), please contact us if you need information about this.

Conditions of use are

- This corpus may be freely used education and non-commercial research.
- It may be redistributed to other academics, students, and researchers
- It should be referred to as the SUMTIME-METEO or SUMTIME corpus. The data providers have asked that they **not** be explicitly mentioned or referred to in papers that use the SUMTIME-METEO corpus (in other words, while we you probably find out who the data provider is from our publications, please do not mention or refer to them in your own papers).
- Published papers should not refer to forecasters or oil rigs by name. We have tried to anonymise or remove this information from the corpus; if you see a name that we have missed, please let us know.

We would also appreciate (but this is not a formal condition of use) if people could let us know what they are doing with the corpus, and about published papers that refer to it.

3. Background

Modern methods of weather forecasting are largely based on computer simulations of numerical weather prediction (NWP) models. These models generate predicted values of various weather parameters such as wind speed, wind direction and precipitation for various time points. In other words, the output of an NWP model is a multivariate time series. Human forecasters use the time series data sets generated by NWP models as the major source of information when writing forecast texts, although they also have access to other information such as satellite weather maps. See section 7 of Sripada et al (2001a) for a more detailed description of the forecasting and writing process.

Weather forecasts are produced for end-users with different information requirements. For instance, forecasts aimed at the general public often describe general outlook of the weather. On the other hand, forecasts aimed at more technically oriented audiences require comparatively more details of the weather. In SUMTIME, we focus on forecasts produced for oil rig staff working offshore in the North Sea. Our collaborating organization works with several NWP models to generate these forecasts. Each NWP model is good at predicting a particular set of weather parameters; this is why several NWP models are used. For instance, the Marine (Wave) model only predicts wind and sea wave data, while the MaxMin (mmo) model predicts general meteorological parameters such as temperature, cloud, and precipitation data.

4. SumTime-Meteo Raw Data

SUMTIME-METEO contains data extracted from three types of files

1. Marine Wave (tab) files – these are the data files generated by the Marine Wave NWP model. This model generates the predicted values for wind and wave related parameters at three hourly intervals.
2. MaxMin (mmo) – these are the data files generated by the MaxMin model. This model generates the predicted values for weather parameters such as cloud, precipitation, at an hourly interval.
3. Forecast (prn) files - the official forecast text, written by human forecasters using the above data

A set of these three files forms one unit of weather data in the sense that they all correspond to one official forecast issued by our collaborating organization. We have been receiving 2 sets of files each day since summer 2000 (with occasional gaps and interruptions); these are for morning and evening forecasts for a particular offshore oil field. SUMTIME-METEO contains data extracted from files received by us between 26 June 2000 and 10 May 2002. We restrict the corpus to this period because after summer 2002 the NWP models changed, and also the forecasts became post-edited versions of computer-generated forecasts (Sripada, Reiter, Hawizy 2005) instead of pure human-written forecasts.

The corpus includes 1119 forecasts (.prn files) (ignoring files which are empty, duplicates, or updates) and similar numbers of data files. Due to gaps we do not have data files for all the forecasts; we only have complete sets (forecast, Marine model, MaxMin model) for 1045 forecasts.

Although we received the raw files as text files, we normally distribute them as a Microsoft Access database. We describe the database’s format below.

Please note that although all times provided to us were in GMT, Access converted them to local times; for example 0600 on 10-April-01 was converted by Access to 0700 (British summer time). This is especially annoying when the clocks change, as then one can get two GMT times which are converted to the same local time; for example 29/10/00 00:00 (GMT) or 29/10/00 01:00 (GMT) both get converted into 29/10/00 01:00 (local time). Amongst other problems, clock changes can also make it harder to calculate things like how fast something is changing. We are not sure what Access will do if the database is opened in a different time zone – please be cautious when using any database field of type Date/Time. If you use CSV dumps of our Access tables, these will contain local UK times, not GMT.

In other words, if your program crashes or acts strangely when handling forecasts in late October (which has happened to us on many occasions), check to see how it handles the shift from summer time to GMT.

4.1. Marine Model Data

The database contains two tables which hold this information

1. WaveTab – index table for Marine Wave data files; this contains one row for each file we received. Columns are
   - filename: Name of the actual text file holding the WaveTab data (primary key)
   - starttime: When the prediction starts; note that most Marine Wave files contain predictions for 3 days, although some contain predictions for 5 days or even a week.
   - forecasttime: The time of the corresponding forecast file (this can be used to join the WaveTab and Forecast tables)
   - ftype: most users should only use files of type “normal”
2. WaveData – This contains the actual NWP data. Each row is the prediction for a particular time from a particular file. Note that each file contains predictions for 20-50 times, and several files may contain predictions for the same time (eg, predictions for 1200 on 10-Apr-01 are provided in 7Apr2001_13.tab, 8Apr2001_02.tab, 8Apr2001_13.tab, 9Apr2001_03.tab, 9Apr2001_13.tab, and 10Apr2001_02.tab). The columns in WaveData are described in the below table

Table 1. Parameters generated by marine model

Column	Weather parameter	Description
filename		Name of data file (this can be used to join with the WaveTab file)
ftime		Time this prediction is for
timetext		Time as specified in the actual data file
winddir	Wind Direction	It is expressed as a string such as ‘N’ representing the direction ‘North’. Numerically ‘N’ corresponds to zero degrees. This model represents wind direction in quanta of 22.5 degrees. Thus wind direction can take any of the possible (360/22.5 =16) 16 values. Enumerating them we have N, NNE, NE, ENE, E, ESE, SE, SSE, S, SSW, SW, WSW, W, WNW, NW, NNW. Wind direction can also be VAR, VR, VRB – this means that the wind direction is variable
windspeed	Wind Speed at 10m height	It is measured in Knots at 10m height. These are also called surface winds.
gust10m	Gust at 10m height	Gusts measured in Knots at 10m height. Note that the forecasters often ignore this information and calculate gust from other parameters
gust50m	Gust at 50m height	Gusts measured in Knots at 10m height. Note that the forecasters often ignore this information and calculate gust from other parameters
sigwave	Significant Wave Height	Expressed in metres.
waveperiod	Wave Period	Expressed in seconds.

swelldir	Swell Direction	It is expressed in a similar way to wind direction, except that CN and CON (“confused”) are used to indicate variable directions
swellheight	Swell Height	Expressed in metres.
swellperiod	Swell Period	Expressed in seconds. This will be blank if swelldir is CON or CN

4.2. MaxMin (mmo) Model File

The database contains two tables which hold this information (similar to 4.1)

3. MMOFiles – index table; this contains one row for each file we received. Columns are
   - filename: Name of the actual text file holding the data (primary key)
   - starttime: When the prediction starts.
   - forecasttime: The time of the corresponding forecast file (this can be used to join the MMOFIles and Forecast tables)
   - ftype: most users should only use files of type “normal”
4. MMOData – This contains the actual NWP data. Each row is the prediction for a particular time from a particular file. Note that each file contains predictions for many times, and several files may contain predictions for the same time (similar to 4.1). Some of the columns in WaveData are described in the below table (contact us if you want information about the other columns

Table 2. Parameters generated by maxmin model

Column	Weather parameter	Description
filename		Name of data file (this can be used to join with the MMOFiles file)
ftime		Time this prediction is for
timetext		Time as specified in the actual data file
clow	Cloud Low	Amount of cloud at lower altitudes measured in octas (0 = no cover, 8 = completely covered)
cmed	Cloud Medium	Amount of cloud at medium altitudes measured in octas
chig	Cloud High	Amount of cloud at high altitudes measured in octas
camnt	Cloud Amount	Amount of the total cloud cover, in octas
Temp	Temperature	Temperature measured in centigrade.
Precip	Precipitation	Precipitation measured in mm/hr
Snow	Snow probability	Probability that precipitation will be snow, as a percentage. Treat numbers less than 0 as if they were 0.
lapse	Lapse rate	The fall/rise of temperature with altitude. Useful in computing the stability of the weather system.

```
1.INFERENCE 0300 GMT, TUESDAY,          25-Dec   2001
LOW 968MB OVER SOUTHERN SWEDEN WILL MOVE EASTWARDS. LOW 976MB
WEST OF BERGEN WILL MOVE SOUTHEASTWARDS TO BE OVER SOUTHERN
DENMARK BY EVENING. A DEPRESSION WILL FORM OFF THE DENMARK STRAIT
AND MOVE SOUTHEAST TO BE OVER THE NORTH OF SCOTLAND BY THURSDAY
AFTERNOON.

2.FORECAST 06-24 GMT, TUESDAY,          25-Dec   2001

=====WARNINGS: WINDS ABOVE 40 KNOTS                         =======

WIND(KTS)      CONF   HIGH
  10M:         N-NNW 28-32 SOON INCREASING 35-40 GUSTS 55, EASING
               25-30 GUSTS 45 LATER
  50M:         N-NNW 35-40 SOON INCREASING 45-50 GUSTS 65, EASING
               30-38 GUSTS 45 LATER
WAVES(M)       CONF   HIGH
  SIG HT:      4.5-5.0 RISING 6.0--7.0, LATER FALLING 5.0-6.0 LATER
  MAX HT:      7.5-8.0 RISING 9.5-11.0, LATER FALLING 8.0-9.5 LATER
  PER(SEC):    8-13
  WEATHER:     SQUALLY WINTRY SHOWERS
  VIS(NM):     OVER 10 REDUCED TO 1-3 IN SHOWERS
  TEMP(C):     3-4 FALLING 1 OR LESS IN SHOWERS
  CLOUD:       3-5 CUSC 1200-2000 BECOMING 5-7 CUCB 600-1000
(OKTAS/FT)     IN SHOWERS
  LIGHTNING RISK : OVER 60 PERCENT IN SHOWERS

3.FORECAST 00-24 GMT, WEDNESDAY,        26-Dec   2001
  WIND(10M):   N-NNW 25-30 GUSTS 45 GRADUALLY EASING NW  8-12
      (50M):   N-NNW 30-38 GUSTS 55 GRADUALLY EASING NW 10-15
  SIG WAVE:    5.0-6.0 FALLING 3.5-4.0
  MAX WAVE:    8.0-9.5 FALLING 5.5-6.5
  WEATHER:     WINTRY SHOWERS GRADUALLY DYING OUT
  VIS:         GOOD EXCEPT IN SHOWERS

4.FORECAST 00-24 GMT, THURSDAY,         27-Dec   2001
  WIND(10M):   NW  8-12 EASING  8 OR LESS FOR A TIME, INCREASING
               NW-N 15-20 LATER
      (50M):   NW 10-15 EASING 10 OR LESS FOR A TIME, INCREASING
               NW-N 18-25 LATER
  SIG WAVE:    3.5-4.0 FALLING 3.0-3.5
  MAX WAVE:    5.5-6.5 FALLING 5.0-5.5
  WEATHER:     SCATTERED WINTRY SHOWERS

5A.LONG RANGE OUTLOOK:  FRI  28-Dec 2001, AND  SAT    29-Dec 2001,
  WIND(10M):   N-NW 15-20 SOON INCRAESING 20-25, EASING 10-15
               SATUDAY MORNING, INCREASING 20-25 LATER
  SIG WAVE:    3.0-3.5 RISING AROUND 5.5, LATER FALLING AROUND 4.0


N.B. THE HIGHEST INDIVIDUAL WAVE THAT MAY BE EXPERIENCED IS OF THE
       ORDER OF TWICE THE SIGNIFICANT WAVE HEIGHT.
```

Figure 4. Official Forecast for the Christmas day of 2001

4.3. Forecast Text File

A sample forecast text is shown in Figure 4. This is as we received it (with the name of the organisation, forecaster, and oilfield removed).

Offshore weather forecasts organise the forecast in terms of forecast periods. Each forecast period describes the forecast for roughly a period of 24 hours. It is common to have three forecast periods or weather forecast information for 72 hours in a SUMTIME-METEO forecast. Before predictions for any individual forecast period are presented, the forecasts contain a description of the general outlook of the weather in that region under the heading 'Inference'. Thus as shown in Figure 4, official forecasts are structured into five sections. The first section is the inference. The next three are the forecasts for the three forecast periods. The fifth one is what is called the long-range forecast. Forecast for each forecast period contains the following elements:

- Wind10M – this part of the forecast text summarises the behaviour of the wind at 10-meter height.
- Wind50M – this part of the forecast text summarises the behaviour of the wind at 50-meter height. Often forecasters write the Wind10M text and then create the Wind50M text by copy-and-edit from Wind10M (or vice-versa)
- Waves Sig. Ht (M) – Significant wave height; average height of the 1/3 highest waves in a record, defined as an approximation to the characteristic wave height (average height of the larger well-formed waves, observed visually). Swell height, swell direction and swell period are also used along with Sig. Ht.
- Waves Max Ht (M) – Maximum wave height for the specified period of time. Swell height, swell direction and swell period are also used along with Max. Ht.
- Wave Period – summarises the wave period data
- Weather – summarises mainly the cloud cover and precipitation.
- Vis – summarises visibility
- Temp – temperature range
- Cloud – summarises amount of cloud

Note that the first five elements of a forecast (wind and wave statements) are mostly produced from the marine wave model data whereas the next four are mostly produced from the MaxMin model data.

Sections are as follows

For AM forecasts (issued in early morning)

- Section 2 covers 0600 to 0000 on issue day
- Section 3 covers 0000 to 0000 on day after issue day
- Section 4 covers 0000 to 0000 on second day after issue day
- Section 5 varies, and is not always present

For PM forecasts (issued around noon)

- Section 2 covers 1500 to 0600. Note that the corresponding Marine WaveTab starts at 1800, not 1500
- Section 3 covers 0600 to 0000 on day after issue day

- Section 4 covers 0000 to 0000 on second day after issue day
- Section 5 varies, and is not always present

The database contains two tables which hold forecast information

5. Forecasts – index table; this contains one row for each file we received. Columns are
   - filename: Name of the actual text file holding the data (primary key)
   - author: The author of the forecast. These are anonymised F1, F2, F3, F4, or F5. This field is blank if we do not know the author
   - forecasttime: The time forecast starts at. This can be used to join with the WaveTab and MMOFiles tables
   - ftype: most users should only use files of type "normal"
   - section1: Inference text (general description of the weather situation)
6. Fields – This contains the specific predictions. Each row contains the prediction for one field in one section of one forecast. Columns are
   - id: unique ID for this field (used in some derived tables, see below)
   - filename: name of forecast file (can be used to join with Forecasts table)
   - section: section of forecast (ie, forecast period)
   - field: field name (eg, "wind10M")
   - ftext: actual text of the field

5. Derived Tables, Queries

In addition to the raw data, the Access database file contains several tables which we derived from the raw data. These include

ParseTuples: We parsed all wind10M statements from Sections 2, 3, and 4. The output of the parser is a set of "tuples", each tuple records the description of the wind at a particular point in time. ParseTuples may be of interest to people who want to relate texts to a conceptual representation instead of raw numerical data. Key columns for this purpose are

- Forecast: the forecast file the text came from
- Section: the section the text came from
- Index: 0 for the first wind descriptor in the text, 1 for the second, etc
- TTime: The time of the wind descriptor. This is computed from the time phrase specified, using a lookup table that maps time phrases to the most common time they represent. A more accurate time can be obtained from windTime table (see below)
- Direction: wind direction, in degrees
- LowSpeed: bottom of wind speed range
- HighSpeed: top of wind speed range
- Gusts: If gusts are mentioned
- Showers: If showers are mentioned
- GustSpeed (at end of table): gust speed (if Gusts are mentioned)

For example, the Section 2 wind10M descriptor from Figure 4,

```
N-NNW 28-32 SOON INCREASING 35-40 GUSTS 55, EASING 25-30 GUSTS
45 LATER
```

is parsed into three tuples, for `"N-NNW 28-32", "SOON INCREASING 35-40 GUSTS 55",` and `"EASING 25-30 GUSTS 45 LATER"`. The actual ParseTuples data for this descriptor is shown below.

Forecast	Section	Index	TTime	Direction	LowSpeed	HighSpeed	Gusts	Showers	GustSpeed
25Dec2001_02.prn	2	0	6	348.75	28	32	No	No	
25Dec2001_02.prn	2	1	9	348.75	35	40	Yes	No	55
25Dec2001_02.prn	2	2	24	348.75	25	30	Yes	No	45

Note that the parser propagates elided information; eg, it assumes that the wind direction remains N-NNW throughout the forecast as no change in wind direction is mentioned.

ParseRejected: This lists all forecast texts which our parser could not parse. Of course ParseTuples will not contain data for these texts.

WindTime: This lists the time of each ParseTuple tuple, calculated by aligning the wind described in the text with the Marine Wave data, as described in Reiter and Sripada (2003) and Reiter, Sripada, Hunter, Yu, Davy (2005). Key columns are

- Forecast, section, index: Identify the tuple in ParseTuples

- MatchTime: The time for this tuple, calculated via alignment with the data file. –99 means this could not be computed.
- MatchQuality: The quality of the alignment (see papers for details). 3 is best, 0 means MatchTime could not be calculated at all.

If you wish to find out the time of a tuple, we suggest you use windTime.MatchTime if this exists and is not –99; otherwise use parseTuples.TTime

Words: This lists all words used in the forecasts (all fields, not just wind10M). Columns are

- ID: This is Fields.id (allows joining with Fields table)
- word: actual word
- pos: index (position) of word in the field

We have also included a few queries in the DB, mainly as examples

- authorwordquery: shows how often each author uses each word
- containsquery: shows all wind10M fields that contain the word "easing"
- matchDataForecast: shows the data files that correspond to each forecast

wind10M: show all wind10M fields from the Fields table

6. Distribution

SUMTIME-METEO is distributed by SIGGEN (http://www.siggen.org/), under the Resources section. The standard SIGGEN distribution is a zip file which contains

- This document (as a PDF file)
- An Access (MDB) database containing the SUMTIME-METEO data
- A Tables subdirectory, which contains CSV (comma-separated-value) text files. This is intended for people who do not have Microsoft Access. There is one CSV file for each table in the database.

7. References

A Belz (2005). Statistical Generation: Three Methods Compared and Evaluated. *Proceedings of ENLG-2005*, pages 15-23

Brown, P; Cocke, J; Della Pietra, S; Della Pietra, V; Jelinek, F; Lafferty, J; Merver, R; and Roossin, P (1990). A Statistical Approach to Machine Translation. *Computational Linguistics* **16**:79-85.

E Reiter and S Sripada (2003). Learning the Meaning and Usage of Time Phrases from a Parallel Text-Data Corpus. In *Proceedings of the HLT-NAACL03 Workshop on Learning Word Meaning from Non-Linguistic Data*, pages 78-85

E Reiter, S Sripada, J Hunter, J Yu, and I Davy (2005). Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence* **67**:137-169

E Reiter, S Sripada, and R Robertson (2003). Acquiring Correct Knowledge for Natural Language Generation. *Journal of Artificial Intelligence Research* **18**:491-516

S Sripada, E Reiter, I Davy, and K Nilssen (2004). Lessons from Deploying NLG Technology for Marine Weather Forecast Text Generation. In *Proceedings of PAIS-2004*

S Sripada, E Reiter, and L Hawizy (2005). Evaluating an NLG System using Post-Edit Data: Lessons Learned. *Proceedings of ENLG-2005*, pages 133-139

S Sripada, E Reiter, J Hunter, and J Yu (2003). Exploiting a Parallel Text-Data Corpus. In *Proceedings of Corpus Linguistics 2003*

S Sripada, E Reiter, J Hunter, J Yu, and I Davy (2001). Modelling the Task of Summarising Time Series Data Using KA Techniques. In A. Macintosh, M. Moulton, and A Preece (eds), *Applications and Innovations in Intelligent Systems IX*, pages 183-196. Springer-Verlag