Exploiting a parallel TEXT - DATA corpus

Somayajulu G. Sripada, Ehud Reiter, Jim Hunter and Jin Yu
Dept. of Computing Science
University of Aberdeen
{ssripada,ereiter,jhunter,jyu}@csd.abdn.ac.uk
Fax: +44 (0)1224 273422

Abstract
In this paper, we describe *SUMTIME-METEO*, a parallel corpus of naturally occurring weather forecast texts and their corresponding forecast data; data that the human authors inspected while writing the forecast texts. We have analysed the corpus to acquire knowledge needed to build a text generator for automatically producing textual weather forecasts from numerical weather prediction data. Although parallel corpora are commonly used for the development and evaluation of machine translation technology, it is fairly novel in the text generation community. Our analyses of the corpus, in some cases, produced ambiguous results that are not useful and reflected inconsistencies in the underlying corpus. Despite the internal inconsistencies, the text-data parallel corpus was helpful in generating initial hypotheses, which were then tested with knowledge from other sources. We also describe how we have used the corpus for evaluating our prototype forecast text generator.

1 Introduction

SUMTIME-METEO is a parallel corpus of 1045 weather forecast texts and the numerical data (output of a numerical weather prediction (NWP) model) that human forecasters examined when writing the texts. While parallel corpora of texts and their translations into other languages are fairly common and are heavily used in machine translation (Brown et al 1990), parallel corpora of data and their "translations" into texts are more unusual. Barzilay and McKeown (2001) report use of parallel translations of texts (i.e. text to text but not data to text) for an NLG application. We are using *SUMTIME-METEO* to understand the structure, content selection rules and linguistic issues in weather forecast generation. More generally, we believe that parallel text-data corpora could be a valuable resource for researchers interested in semantics, as they provide empirical data on how non-linguistic content is expressed as words and sentences. *SUMTIME-METEO* is a naturally occurring corpus in the sense that the forecast texts were written for actual clients, and were not artificially written just for our project.

The corpus is built as part of the *SUMTIME* project (http://www.csd.abdn.ac.in/research/sumtime). In the project our objective is to develop generic techniques for summarising time series data. Towards this end, we are working in three different domains – meteorology, gas turbines and neonatal intensive care units. In the domain of meteorology, we are trying to build a system, which automatically generates weather forecasts from numerical data, in collaboration with WNI Oceanroutes, Aberdeen, U.K. Our weather forecast generator, ***SUMTIME-MOUSAM*** is based on knowledge acquired from several sources, including analysis of the *SUMTIME-METEO* corpus. It is currently installed in WNI and is being run in parallel with human forecasters (with computer-generated forecasts compared against manually written ones), and may soon go "live" for some specialised types of forecasts.

2 Corpus description

2.1 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 company staff supporting offshore activities in the North Sea. Our collaborating organization works with three NWP models to generate these forecasts - Marine model, MaxMin model and Media Kernel model. Each NWP model is good at

predicting a particular set of weather parameters; this is why several NWP models are used. For instance, Marine model is good at predicting wind and sea wave data where as the other two (which have many overlapping parameters) are good at predicting other parameters such as precipitation, temperature etc.

2.2 Corpus contents

We receive weather data corresponding to a specific offshore oil-drilling site in the form of three files corresponding to the three NWP models and one file containing the manually written forecast text:

1. Files with extension .tab (or .etb) – these are the data files generated by the Marine model. This model generates the predicted values for wind and wave related parameters at three hourly intervals.
2. Files with extension .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. Files with extension .csv – these files are generated by Media Kernel model and contain roughly similar information to the .mmo files.
4. Files with extension .prn - the official forecast texts issued by our collaborating organization (written by human forecasters using the above data)

A set of these four files forms one unit in the corpus because they all correspond to one official forecast issued by our collaborating organization. WNI has been sending us two such sets a day since the summer of 2000 (with occasional gaps and interruptions); most of these files are the 0300 and 1200 forecasts for the Magnus, Thistle, and NW Hutton oil fields. The current distribution version of our corpus contains files received by 10 May 2002. After this date our prototype system was installed at WNI and therefore we assumed that the human authors might write texts under its influence. This 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; for example, we only have the corresponding .tab file for 1045 forecasts.

2.3 Marine model file

These files consist of a header and a body as shown in Figure 1.

Header: Essentially the header information associates the data in the body part to a specific location and time. The first line of the tab (etb) file states information about the location (the name(s) of the location) to which the predicted data belongs. The sample data shown in Figure 1 belongs to three locations east of Shetland – MAGNUS, THISTLE and NW HUTTON. It should be noted that NWP models do not work with actual location names. Instead they compute data for the entire surface of the earth by dividing it into a two dimensional array called 'grid' and output data for a 'grid point' (node in the array). A grid point might map onto more than one location name. The second line of the header states the start date of the predicted data in GMT. The format of the date string is "dd-mm-yy". In Figure 1, the date is 24 October 2000.

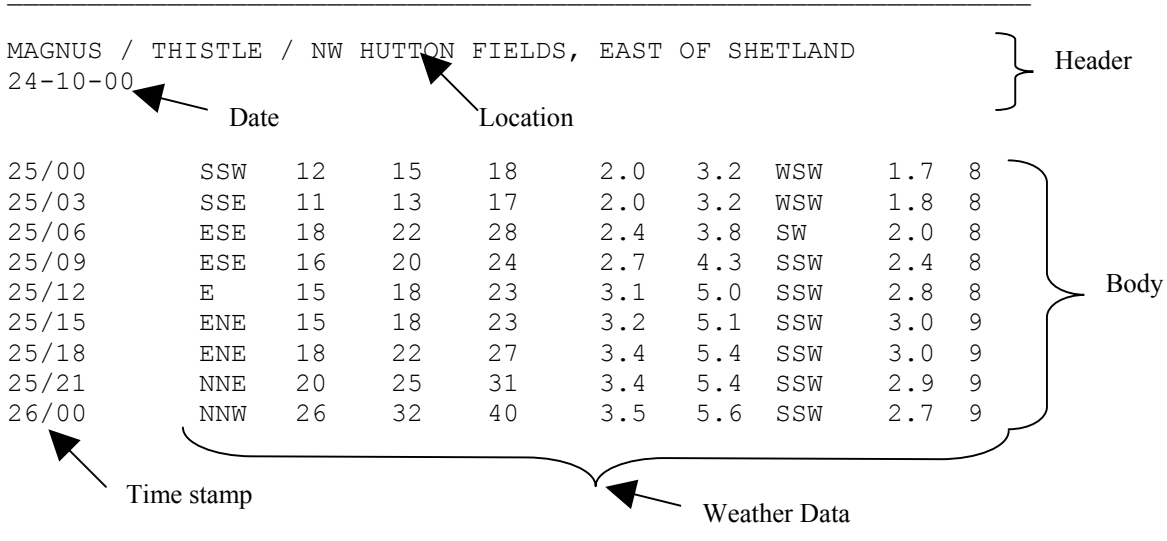

Figure 1. A portion of .tab (.etb) file for 24-10-2000.

Body: This part of the file contains the weather data in rows and columns. Each row holds weather data predicted for a point of time (time stamp). The first column holds the time stamp and each of the other columns holds the value of a specific weather parameter. The format of the time stamp is "dd/HH". In Figure 1, the time stamp in the first row is '25/00', which means that the data in that row belongs to 0000 hours on 25-10-00. The data columns are described below in Table 1 (the first column is already described as time stamp):

Column	Weather parameter	Description
2	Wind Direction	It is expressed as a string such as 'E' in Figure 1 representing the direction 'East'. Numerically 'E' corresponds to 90 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.
3	Wind Speed at 10m height	It is measured in Knots at 10m height. These are also called surface winds. In Figure 1, for the time stamp 25/06 it is 18 Knots.
4	Gust at 10m height	Gusts measured in Knots at 10m height. In Figure 1, for the time stamp 25/06 it is 28 Knots.
5	Gust at 50m height	Gusts measured in Knots at 50m height. In Figure 1, for the time stamp 25/06 it is 28 Knots.
6	Significant Wave Height	Expressed in metres. In Figure 1, for the time stamp 25/06 it is 2.4
7	Wave Period	Expressed in seconds. In Figure 1, for the time stamp 25/06 it is 3.8
8	Swell Direction	It is expressed in a similar way to wind direction. In Figure 1, for the time stamp 25/06 it is SW.
9	Swell Height	Expressed in metres. In Figure 1, for the time stamp 25/06 it is 2.0
10	Swell Period	Expressed in seconds. In Figure 1, for the time stamp 25/06 it is 8.

Table 1. Parameters generated by marine model

```
_____________________________________________________________________

OCEANROUTES SPECTRAL WAVE AND WEATHER FORECAST.
DUTY FORECASTER AVAILABLE AT ALL TIMES.PHONE ABERDEEN (01224) 248080
FORECAST FOR:-

MAGNUS, THISTLE AND NW HUTTON FIELDS, EAST OF SHETLAND

1.INFERENCE 0300 GMT, TUESDAY,          24-Oct   2000
LOW PRESSURE WILL MOVE SLOWLY AWAY NE INTO THE NORWEGIAN SEA, WITH
A RIDGE MOVING EAST ACROSS THE NORTH SEA THIS EVENING. A VIGOROUS
FRONT WILL CROSS THE CENTRAL NORTH SEA TONIGHT, WITH THE PARENT
LOW MOVING EAST ACROSS NORTHERN SCOTLAND TOMORROW MORNING THEN
CONTINUING EAST ACROSS THE NORTHERN NORTH SEA BY TOMORROW NIGHT.
                                                AUTHOR NAME

2.FORECAST 06-24 GMT, TUESDAY,          24-Oct   2000

=====WARNINGS: NIL                                              =======

WIND(KTS)       CONFIDENCE HIGH
  10M:          SW 10-14 VEERING W 18-22 GUSTS 30 BY LATE MORNING
                THEN BACKING AND EASING SSW 12-16 THIS EVENING
  50M:          SW 12-18 VEERING W 23-28 GUSTS 38 BY LATE MORNING
                THEN BACKING AND EASING SSW 15-20 THIS EVENING
```

```
WAVES(M)        CONFIDENCE HIGH
  SIG HT:       2.5-3.0 FALLING 2.0-2.5
  MAX HT:       4.0-5.0 FALLING 3.0-4.0
  PER(SEC):     SEA 3-4 RISING 5 FOR A TIME, SWELL 8-9
  WEATHER:      SHOWERS DYING OUT THIS EVENING
  VIS(NM):      OVER 10 FALLING 3-5 IN SHOWERS
  TEMP(C):      10-11
  CLOUD:        5-7 CU/SC 1500-2500 WITH 7-8 CU/CB 1000 IN SHOWERS
(OKTAS/FT)
LIGHTNING RISK:MODERATE (50-60 PER CENT) FALLING NIL LATER

3.FORECAST 00-24 GMT, WEDNESDAY,        25-Oct   2000
  WIND(10M):    SSW 12-16 BACKING ESE 16-20 IN THE MORNING, BACKING
                NE EARLY AFTERNOON THEN NNW 24-28 LATE EVENING
       (50M):   SSW 15-20 BACKING ESE 20-25 IN THE MORNING, BACKING
                NE EARLY AFTERNOON THEN NNW 30-35 LATE EVENING
  SIG WAVE:     2.0-2.5 RISING 3.0-3.5 BY AFTERNOON
  MAX WAVE:     3.0-4.0 RISING 5.0-5.5 BY AFTERNOON
  WEATHER:      RAIN SOON, CLEARING TO SHOWERS IN THE EVENING
  VIS:          GOOD BECOMING MODERATE IN RAIN

4.FORECAST 00-24 GMT, THURSDAY,         26-Oct   2000
  WIND(10M):    NNW 24-28 EASING 14-18 IN THE MORNING THEN BACKING
                NW 10 OR LESS IN THE EVENING
       (50M):   NNW 30-35 EASING 18-23 IN THE MORNING THEN BACKING
                NW 12 OR LESS IN THE EVENING
  SIG WAVE:     3.0-3.5 FALLING STEADILY TO 1.5-2.0 LATER
  MAX WAVE:     5.0-5.5 FALLING STEADILY TO 2.5-3.0 LATER
  WEATHER:      SHOWERS DYING OUT

5A.LONG RANGE OUTLOOK:  FRI  27-Oct 2000, AND  SAT    28-Oct 2000,
  WIND(10M):    NW 10 BACKING S-SE 22-26 BY MID FRIDAY, EASING 18-22
                BY EARLY SATURDAY THEN RISING SSW 35-40 LATER
  SIG WAVE:     1.5-2.0 RISING 2.0-3.0 ON FRIDAY AND 5.0-5.5
                LATER ON SATURDAY

5B.LONG RANGE OUTLOOK:  SUN  29-Oct 2000, AND  MON    30-Oct 2000,
  WIND(10M):    S-SW 35-40 EASING 10-14 ON SUNDAY, RISING S 20-24
                ON MONDAY
  SIG WAVE:     5.0-5.5 FALLING 3.0-4.0 BY LATE SUNDAY

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

Figure 2. WNI Forecast for 24-10- 2000

2.4 Forecast text file

A sample forecast text is shown in Figure 2.

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 an official forecast issued by WNI. Sometimes as in Figure 2, there is also a long-range forecast. Before predictions for any individual forecast period are presented, the official forecasts issued by WNI contain a description of the general outlook of the weather in that region under the heading ‘Inference’. Thus as shown in Figure 2, 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. This statement is based on parameters, wind speed, wind direction and gust at 10 meter from the marine model as shown in Figure 1.
Wind50M – this part of the forecast text summarises the behaviour of the wind at 50-meter height. Wind direction and Gust50m are also used along with Wind50m for writing this element.
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

Forecast texts in the first five fields (Wind10M, Wind50M, Sig. Wave, Max. Wave and Wave Period fields of Figure 2) are all produced primarily from the data shown in Figure 1. For example, the Wind10m field for the forecast period 06-24 GMT 25 Oct 2000 as shown in Figure 2 summarises data from the second (Wind direction), third (Wind speed at 10 meter altitude) and fourth (gust at 10 meter altitude) columns of the marine model data shown in Figure 1. The last four fields (Weather, Visibility, Temperature and Cloud) of the forecast shown in Figure 2 are produced primarily using the data from the other two models (MaxMin and Media Kernel). Details about the contents of each of these models have been presented in (Sripada et al 2002a).

2.5 Distribution of the corpus

SUMTIME-METEO can sometimes be distributed to other researchers. Interested researchers should contact Dr. Ehud Reiter (email: ereiter@csd.abdn.ac.uk).

3 Knowledge acquisition studies using SUMTIME-METEO

In this section, we describe how we are using the parallel text–data corpus in the **SUMTIME** project to acquire knowledge to build a weather forecast text generator, **SUMTIME-MOUSAM**. In our work, we follow the three-stage reference architecture for text generation as described in (Reiter and Dale 2000). Our main objective for corpus analysis has been to acquire knowledge needed for all the three stages of text generation (document planning, micro-planning and realization) from the parallel corpus. Corpus analysis happens to be one of the techniques in our knowledge acquisition studies. We have also tried other techniques familiar in the expert system community (Scott et al 1991). Our idea has been to initially collect knowledge from multiple sources and to finally consolidate the different findings into a consistent and usable model to be used for building **SUMTIME-MOUSAM**.

For our discussion here we focus on texts from the Wind10M field (please refer to the Wind10M field in Figure 2) although our software generates all the fields. Forecasters write these texts primarily by inspecting data from the second (Wind direction) and the third (Wind speed at 10 meters altitude) columns of the marine model output (please refer to the Figure 1). Other data that is included in Wind10M texts such as gusts is secondary and therefore omitted in this discussion.

3.1 Content determination

The first task of a data summarization system is to select data items from the input data set to include in the summary text. Learning the content determination rules automatically, from the parallel corpus, was hard. Manual analysis of the parallel corpus revealed that human forecasters select data points that are representative of trends in the input data set. Accordingly we have used a well-known segmentation algorithm from the KDD community known as bottom-up segmentation algorithm (Keogh et al 2002) to segment the input data set (for example, Wind speed and direction at 10 meter altitude separately) in order to determine the trends in the input. However, our 'think aloud' sessions with an expert forecaster where he spoke out the process while writing forecast texts revealed an alternative procedure for data selection which is based on identifying 'significant' changes in the input data.

Time	Wind Speed in Knots	Wind Direction
10-10-2003 00:00:00	5	S
10-10-2003 03:00:00	6	S
10-10-2003 06:00:00	7	S
10-10-2003 09:00:00	8	S
10-10-2003 12:00:00	9	S
10-10-2003 15:00:00	10	S
10-10-2003 18:00:00	11	S
10-10-2003 21:00:00	12	S
11-10-2003 00:00:00	13	S

Table 2. Fictitious Wind Speed Data for 10-10-2003

For example, consider the fictitious wind data shown in Table 2. The wind speed in this example case is monotonically increasing from 5 knots to 13 knots and the wind direction has always been S, southern. Segmentation approach fits the entire data set into one single rising line and accordingly selects the first (data corresponding to the time 00 hours on 10-10-2003) the last (data corresponding to the time 00 hours on 11-10-2003) data points. The wind statement generated in this case could be something like '*S less than 10 rising steadily 10 –15 by midnight*'. On the other hand, expert approach is based on an externally controlled threshold value for deciding 'significant' changes. If we assume a value of five (which is what the expert suggested) for the threshold, the expert approach selects data points as explained next. After picking up the first data point, this approach looks for the data point that is greater or equal (also smaller or equal) than the first data point by five (the threshold). In the example case, this happens to be the data corresponding to the time 15 hours on 10-10-2003. The wind statement generated in this case could be something like '*S less than 10 rising steadily 8-13 by mid afternoon*'.

We have implemented both these procedures in our system and compared the two using an intrinsic evaluation (Sripada et al 2002b). The evaluation results indicated that segmentation procedure acquired from corpus studies is better than the procedure suggested by the expert. The expert meteorologist with whom we did the think aloud session also agreed with the result of evaluation eventually. This can be explained by observing that the meteorologist not being an expert in designing computer algorithms initially failed to develop a model reflecting his selection process accurately.

3.2 Learning the meaning and usage of time phrases and change verbs

The next task in a data summarization system is to make appropriate lexical choices to express the selected content. Here, we need good models of word meaning (or usage). Although there are existing procedures from the field of linguistics and lexicography for acquiring meanings of words, a parallel corpus with its text to data associations offers a novel resource to link words to numerical data and thereby to explore their meaning (or usage) in terms of the numerical data. Preliminary analysis of corpus showed that the majority of forecast texts (please refer to the forecast in Figure 2) are made up of time phrases (e.g. '*by evening*' and '*in the morning*') and change verbs (e.g. '*veering*' and '*backing*'). Thus we focused our attention on learning the meaning of these two types of phrases.

First we describe our analysis of the parallel corpus for learning the interpretation of time phrases. Particularly, our objective is to determine what time the forecaster meant when he used a time phrase such as '*by evening*'. This knowledge will help our forecast generator decide the specific time phrase to mark a change in wind speed, say at 1800 hours. It should be mentioned in this context that expressions of time do not get standardised in the same way as expressions of cloud cover or precipitation. Therefore, forecasters don't have access to a list of acceptable time phases and their time mappings. Also it is very rare for forecasters to mention numerical time values in forecast texts. For this study, we aligned phrases from forecast texts with numerical data from the numerical weather simulation, and used this alignment to infer the meanings of time phrases. The precise alignment process is explained next.

Sr. No.	Phrases	Information			
		Speed	Direction	Time Phrase	Change Verb
1	SSW 12-16	12-16	SSW	None	None
2	BACKING ESE 16-20 IN THE MORNING,	16-20	ESE	IN THE MORNING	BACKING
3	BACKING NE EARLY AFTERNOON	16-20	NE	EARLY AFTERNOON	BACKING
4	THEN NNW 24-28 LATE EVENING	24-28	NNW	LATE EVENING	BACKING

Table 3. Parser Output for Wind10M text for 25-10-00.

As stated earlier we performed our analysis on forecast texts from the Wind10m field. As shown in Figure 2 there are three Wind10m texts per forecast corresponding to the three days for which the forecast is issued. Initially we parsed these texts with a simple parser tuned to the linguistic structure of these texts. The parser first breaks each Wind10M statement up into phrases and then extracts information such as wind speed and direction from each phrase. This extracted information is recorded into a database. For example, for the Wind10m text from the 06-24 GMT, 25 Oct 2000 forecast in Figure 2, the parser extracts phrases and records information as shown in Table 3.

In Table 3, phrase 3 has the speed information missing and phrase 4 has the verb missing. Our parser works in such cases by carrying forward the values from the previous phrase. Thus the wind speed for the third phrase is same as the second one and the change verb for the fourth phrase is same as the third. We have noticed that such cases of elision (omitted expressions) are very common in forecast texts, and carried out a separate study to acquire rules for eliding words or phrases. We describe this study in section 3.3. Our parser produced 8198 wind phrases and their related information from all the Wind10M statements in our corpus. For learning the meaning of time phrases not all of the 8198 phrases are useful for various reasons. For instance, the first phrase in Table 2 does not have a time phrase and therefore of no use in the corpus analysis (in fact, less than 1% of the initial phrases in wind statements included a time phrase). After applying all such restrictions, the useful number of time phrases reduced to 3654.

Time	Most common phrase in corpus	Expert suggested phrases	Phrases used in **SUMTIME-MOUSAM**
0	By late evening	Around midnight	By midnight
3	Tonight	In early hours	After midnight
6	Overnight	In early morning	By early morning
9	By midday	During midday	By morning
12	By midday	Around midday	By midday
15	By mid afternoon	In mid afternoon	By mid afternoon
18	By evening	In early evening	By early evening
21	By evening	During night	By evening

Table 4. Time Phrase Mappings from Corpus studies, Experts and **SUMTIME-MOUSAM**

Next we associated each wind phrase with an entry in the corresponding data file. This process, known as alignment, is very important for performing analysis on a parallel corpus (Och and Ney 2000). Good alignment techniques produce better results in the subsequent corpus analysis. Initially we have aligned all those phrases that matched with an entry in the data file without ambiguity. The proportion of such aligned pairs was 43%. Then we have used simple heuristics to resolve ambiguities in the remaining cases. After applying our heuristics we have been able to find 2539 aligned pairs, which is about 70% of the collected phrases. We are still working on improving the alignment process. More details about this study can be found in (Reiter and Sripada 2002). Using the aligned pairs (wind phrase and data entry) we have then mapped all the time phrases to their corresponding interpretations. For example, if the text contains the phrase "VEERING SW 10-14 BY EVENING" and the 1800 hours entry of the NWP model output is the only one with a direction of SW and a speed in the 10-14 range, then we assume that in this instance "BY EVENING" means 1800. This analysis showed that time phrases such as 'by evening' had much vaguer and more varied time mappings than we expected. This is bad for us because we cannot use these mappings in our forecast text generator. We have approached the experts

to verify the results of our corpus analysis and to suggest time phrase interpretations. The first column in Table 4 shows the various times for which we need the time phrases. The second and third columns show the time phrases discovered in our corpus analysis and those suggested by the experts. The last column shows the consolidated time phrases that we finally used in **SUMTIME-MOUSAM**.

We have also used the corpus for learning what verbs to use to describe changes in wind speed and direction. For example, the Wind10M text for 26-10-2000 in Figure 2 uses verbs 'easing' and 'backing'. Our objective in this study is to learn which change in the underlying data prompted the forecaster for using these verbs. This was done by applying the machine learning algorithm, Ripper (Cohen 1995) to our corpus. This study highlighted that there were substantial individual differences in the use of near-synonyms such as EASING and DECREASING. Predictive rules could be created for individual forecasters, but not for the corpus as a whole (if we did not allow the rules to include author). We also ran C4.5 (Quinlan 1993) on our corpus; this gave similar results.

3.3 Learning rules of expression

The next important task in data summarization is to learn when to elide (omit) words and phrases. We have used only the forecast texts for this study without the parallel data component of the corpus. The results of this analysis were surprising because they disagreed with what expert forecasters had told us. For example, we noticed from our corpus studies that when the wind speed (or direction) varied steadily throughout a forecast period, (such as in fictitious data set shown in Table 2) forecasters often omitted a time phrase. For the example data in Table 2, the forecasters might something like '*S less than 10 rising steadily 10 –15'* rather than '*S less than 10 rising steadily 10-15 by midnight*'. In contrast, the experts felt that end users find it better when a time phrase is included because in this case they are not required to remember when the forecast period ends. We suspect this may reflect forecasters trying to use latest data available to them; in other words, the corpus-derived rules in this case may *not* be appropriate ones to include in a computer system.

We have carried out a number of other analyses on the corpus to understand how forecast content is expressed in the texts. A number of expression level variations have been observed in the textual forecasts. For example, single digit numerical values have been expressed with and without leading zero such as '*06*' and '*6*'.

3.4 Evaluation of prototype forecast text generator

We have used the corpus for evaluating the output of our forecast generator by comparing computer-generated forecasts to manually written ones. This was done at a conceptual level that captures the content in a forecast at a higher level independent of the variations in expression. Comparison at the conceptual level was chosen because text level comparison would be difficult with all the expression level variations discovered in the above studies. For extracting the conceptual structures from the corpus texts we have used a parser, which is an improvement over the one described in section 3.2. It extracts additional information from the wind phrases. Instead of writing the extracted information into a database the parser here populates the conceptual structures. Conceptual representation of wind forecast consists of a tuple with the following elements

- Time
- Wind speed lower range
- Wind speed upper range
- Wind direction (in degrees)
- Modifiers (gusts, shower, steady-change, gradual-change)

For example, for the Wind10m text from the 06-24 GMT, 25 Oct 2000 forecast in Figure 2, we have four tuples as shown below:

(00, 12, 16, 202.5, none)
(06, 16, 20, 112.5, none)
(15, 16, 20, 45, none)
(00, 24, 28, 337.5, none)

For this study, we have used one version of our prototype forecast text generator to produce the forecast in the conceptual form starting from the NWP model data in our parallel corpus. Using our

parser we have generated the conceptual forecast from the human-written forecast text for the same data. We have compared both the forecasts in their conceptual forms and collected all those cases in the corpus where the machine-generated forecast is substantially different from the human written forecast. We have externally controlled the matching process by defining different matching criteria. In a majority of cases (about 85%) there was a good match between the two. Upon careful observation we have realised that these cases correspond to days when there was nothing complex happening with the weather, they involve routine forecasting. Though low in number, a majority of the other cases (15%) represent days when more complex weather conditions existed. We believe that the human forecasters used an overview of the weather while writing these (15%) texts as described in (Sripada et al 2001b).

4 Conclusion

We have described how we acquired knowledge from a parallel text-data corpus of naturally occurring weather forecasts for building a prototype forecast text generator, *SUMTIME-MOUSAM*. The knowledge acquired from the corpus was useful in a few cases but in many other cases required clarifications with expert meteorologists. Particularly, unlike what we expected, the corpus gave ambiguous results when we tried to find what words (or phrases) mean and what rules govern their expression in forecast texts. Among other things, this might partly be because the human forecasters who wrote these texts were writing many forecasts in a day under time pressure and therefore failed to maintain consistency. Despite these problems we found the parallel corpus useful and plan to explore it further. Particularly, we plan to try alignment techniques from the machine translation community, which might help us in obtaining better results.

Acknowledgements

Many thanks to our collaborators at WNI/Oceanroutes especially Ian Davy and Dave Selway; this work would not be possible without them! This project is supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant GR/M76881.

References

Barzilay R, McKeown K 2001 Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL-01)*, pages 50-57.

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.

Cohen W 1995 Fast Effective Rule Induction. San Francisco, Morgan Kaufmann.

Keogh E, Chu S, Hart D, Pazzani M 2001 An Online Algorithm for Segmenting Time Series. In *Proceedings of IEEE International Conference on Data Mining*. pp 289-296.

Och F, Ney H 2000 A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000),* pages 1086-1090.

Quinlan R 1993 C4.5: Programs for Machine Learning. San Mateo, CA, Morgan Kaufmann.

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

Reiter E, Sripada S 2002 Should Corpora Texts be Gold Standards for NLG? In *Proceedings of INLG-2002*, pages 97-104.

Scott A, Clayton J, Gibson E 1991 *A Practical Guide to Knowledge Acquisition*. Addison-Wesley.

Sripada S, Reiter E, Hunter J, Yu J 2001a SUMTIME: Observations from KA for Weather Domain. Technical Report AUCS/TR0102, Department of Computing Science, University of Aberdeen.

Sripada S, Reiter E, Hunter J, Yu J 2001b A Two-stage Model for Content Determination. In *Proceedings of the 8th ACL-EWNLG'2001,* pp3-10.

Sripada S, Reiter E, Hunter J, Yu J 2002a SUMTIME-METEO: Parallel Corpus of Naturally Occurring Forecast Texts and Weather Data. Technical Report AUCS/TR0201, Department of Computing Science, University of Aberdeen.

Sripada S, Reiter E, Hunter J, Yu J 2002b Segmenting Time Series for Weather Forecasting. In: Macintosh, A., Ellis, R. and Coenen, F. (ed) *Proceedings of ES2002*, pp. 193-206.