SYNTHESIZING WEATHER FORECASTS FROM FORMATTED DATA

R.Kittredge and A.Polguère
Département de Linguistique
Université de Montréal

E.Goldberg
Atmospheric Environment Service
Environment Canada, Toronto

Abstract

This paper describes a system (RAREAS) which synthesizes marine weather forecasts directly from formatted weather data. Such synthesis appears feasible in certain natural sublanguages with stereotyped text structure. RAREAS draws on several kinds of linguistic and non-linguistic knowledge and mirrors a forecaster's apparent tendency to ascribe less precise temporal adverbs to more remote meteorological events. The approach can easily be adapted to synthesize bilingual or multi-lingual texts.

1. Natural Language Report Synthesis

We use the term "natural language report synthesis" (NLRS) to describe the process of creating well-formed text which summarizes formatted data in a given domain using a style which mirrors the conventions of professional report writers for that domain.

NLRS for highly restricted domains was first demonstrated in the work of Kukich (1983) on "knowledge-based generation" of stock market reports. Kukich's ANA system produces professional-sounding stock market summaries using a daily trace of Dow Jones' half-hourly quotations for the market average and major indices. Both ANA and the analogous FRANA system for French (Contant 1986) have used a phrasal lexicon approach (Becker 1975) which limits the generality of the linguistic component, but which seems to suffice for small and stereotyped domains. The work described below represents a more modular approach to NLRS as well as a new application domain.

2. Synthesis of Arctic Marine Weather Forecasts

The RAREAS system was developed during a five-month effort to explore the feasibility of synthesizing marine weather bulletins from formatted weather forecast data. The particular task was to produce Arctic marine forecasts for five forecast areas to the east of Baffin Island (known as FPCN25 forecasts). Marine forecasts are one of several types of weather bulletin based on the same basic weather data, each type emphasizing the conditions of interest to a particular community of users. In the case of marine bulletins, linguistic emphasis is placed on wind direction and speed, dangerous wind and freezing spray conditions, etc. RAREAS is designed to be sufficiently modular and flexible so as to allow easy extension and adaptation to other types of weather bulletin (e.g., agricultural bulletins, public weather forecasts). Although the current project seems to have proved the feasibility of automatically synthesizing weather forecasts, extensive testing and refinement is required before RAREAS or any successor can be introduced into daily use.

The RAREAS system is the natural language component of the MARWORDS project, which envisages automating the process of creating bulletins from meteorological information. In the current manual procedure all the available meteorological information (observations, radar and satellite imagery, and numerical weather prediction products) is made available to the weather forecaster. The weather forecaster must correctly diagnose the meteorological processes which will affect his particular area of interest throughout the forecast period, and then translate this knowledge into appropriate textual forecasts for various users.

In the proposed automated process, MARWORDS will use predicted values for meteorological parameters such as wind speed and direction, cloud cover, and others. In some cases, these predictions could be obtained directly from numerical weather prediction products. In most cases though, they would still be the result of a manual (i.e., human) forecasting procedure. MARWORDS will significantly reduce the workload on the forecaster, making it possible to focus more attention on meteorological problems.

In the normal course of events, the predicted values make up a continuum in both time and space. For simplicity, values are often given at regular steps in time (e.g., hourly) and space (either at grid points, or at weather observing sites). Alternatively, forecast parameters may be given in terms of significant changes only. MARWORDS is flexible enough to accept both types of data description. In fact, the structure and nature of the required data is a problem which needs more work to resolve.

3. Design of the RAREAS system

A major task in designing RAREAS was the definition of an input data format which properly divides the work between the MARWORDS expert system, which computes predicted values of weather parameters based on large-scale observations, and RAREAS itself, which interprets that data under local conditions for the purpose of marine forecasts. The format and its permissible content should be sufficiently rich in expressive power to reflect the nuances found in natural language forecasts. Ideally, the expert system should be kept as independent of forecast purpose as possible. RAREAS should therefore take care of all matters related to subjective evaluation of the data (e.g., importance of individual parameters of marine forecasts), as well as the linguistic expression of data values and data relations.

In its current implementation RAREAS reads the formatted forecast data and carries out (sequentially) the following major operations:

- reading and parsing of formatted input data, with the interpretation of certain coded values;
- checking of data for consistency and plausibility, using databases of geographical and meteorological information;
- insertion of default values when needed;
- detection of conditions which are hazardous for marine operations (e.g., freezing spray, calculated as a function of forecast wind speed and air temperature, and of a seasonally and regionally adjusted water temperature taken from the database);
- "merging of areas", namely, a check for similarity in the data for contiguous forecast areas; when similarity threshold conditions are satisfied a single report formula is created for the merged areas under a header which lists those areas;

- suppression of data not sufficiently salient for explicit inclusion in the report (e.g., temperature is generally dropped after its use to check if freezing spray conditions are present);
- synthesis of pre-linguistic ("logical") representation for each sequence of weather events;
- interpretation of transitions between weather events into same pre-linguistic form;
- segmentation of logical structures into more independent pre-linguistic clauses and sentences;
- mapping of clausal form into English word strings, using proper terminology and style.

These diverse functions are carried out by relatively independent modules written in MProlog.

4. A Sample Report

The following simplified example (figure 1) shows the input formatted data, using mnemonic descriptors, for the Frobisher Bay forecast area. (Here, we leave aside the problem of possibly merging reports from the seven areas or sub-areas that are considered together).

```
0400 mon 85/10/16 end.

frob wind 0 10 &
            nt 6 dir 140 speed 15 &
            nt 9 dir 90 speed 30 &
            nt 6 dir 50 speed 35 &
            nt 12 dir 20 speed 40
            sky bkn
            wea snow &
               fog per &
               mist per
            temp -1
            end.
```

Figure 1. Sample RAREAS formatted input.

The formatted data identifies the Greenwich time of report validity, the date and area concerned, and then specifies initial values for each important weather parameter. Subsequent changes in the value of a parameter are preceded by the number of hours until the forecast change. Localized exceptions to the general forecast are preceded by a coded sub-area specification. At present, input data is limited to the six most important parameters: (1) wind direction, (2) wind speed, (3) cloud cover classification, (4) precipitation types (if any), (5) precipitation frequency and intensity rating, and (6) air temperature. Further forecast parameters which are functions of the input parameters (e.g., warnings and visibility ratings) are calculated by the first non-linguistic module.

After reading and analysis, the data is manipulated in clausal form through data checking, area unification and data suppression stages mentioned above. It is then translated into a "logical form" just before input to the linguistic modules.

Linguistic modules first calculate the values of signficant semantic features of incipient lexical items, particularly regarding direction and degree of changes. For example, winds which change direction in a clockwise direction will be described lexically as "veering" to the new direction, whereas winds which change in a counter-clockwise direction are described as "backing". Initial lexical instantiation uses the most precise term available in the lexicon. Subsequent segmentation into sentences may juxtapose clauses in such a way that lexical variation is desirable. Precise terms may then be replaced by synonymic variants, or by more general (hyperonymic) lexemes.

Figure 2 gives the final textual form of the marine forecast corresponding to the data of figure 1 above.

```
MARINE FORECASTS FOR ARCTIC WATERS ISSUED BY
ENVIRONMENT CANADA AT 9:00 PM MDT MONDAY 16
OCTOBER 1985.
VALID UNTIL MIDNIGHT TUESDAY WITH AN OUTLOOK
FOR WEDNESDAY.

FROBISHER-BAY
GALE WARNING ISSUED ...
WINDS LIGHT BECOMING SOUTHEASTERLY 15 EARLY
TUESDAY MORNING THEN BACKING AND STRENGTHEN-
ING TO EASTERLY 30 TUESDAY AFTERNOON THEN
STRENGTHENING TO NORTHEASTERLY GALES 35 TUES-
DAY EVENING. MOSTLY CLOUDY WITH SNOW. FOG AND
MIST PATCHES. VISIBILITY FAIR IN SNOW, FAIR IN MIST
AND POOR IN FOG.
OUTLOOK FOR WEDNESDAY ... GALE FORCE
NORTHEASTERLIES BECOMING GALE FORCE NORTHER-
LIES.
```

Figure 2. RAREAS output for data of fig. 1 above.

5.Knowledge Sources for Report Synthesis

The RAREAS architecture isolates different types of linguistic and non-linguistic knowledge within appropriate modules. Our grammatical, lexical, rhetorical and stylistic description is based on an examination of all the marine bulletins (manually) produced for the FPCN25 region during the 1983 and 1985 seasons (some 50,000 words in all).

Examination of this extensive corpus of English has led to a fairly detailed grammar of this sublanguage (cf. Harris 1968, Kittredge and Lehrberger 1982).

Linguistic knowledge is broken down into several types:

- lexical semantics, including conditions for appropriate usage of words in terms of corresponding data values, and frequency preferences among synonymous terms in the sublanguage of marine bulletins;
- syntactic patterns, including the possible and preferred sentence patterns for expressing messages of given types; a second type of syntactic knowledge concerns the rules for deleting repeated sentence constituents when two or more propositions are fused into a single report sentence;
- simple principles of text organization, specific to the variety of text to be synthesized, and hence a function of the data salience hierarchy (see below);

Non-linguistic knowledge is of three types:

- geographical knowledge for each forecast area including (1) its time zone, (2) its limits of latitude and longitude, and (3) the names of adjoining areas (to allow recursively merging adjacent areas in case of similar meteorological regimes);
- meteorological data including (1) mean temperature values for air and water during each month of the Arctic shipping season (June through October) and (2) record values for temperature & wind speed;
- an "archive" of data from preceding reports, used to verify if dangerous wind warnings or freezing spray warnings are in effect.

Geographic knowledge is used primarily during the attempt to merge reports for adjoining areas. However time zone data is used to calculate local time associated with meteorological phenomena, and hence allow attribution of appropriate temporal descriptors (e.g., "by late afternoon"). Input data to the system has only the Greenwich reference time used by meteorologists.

6. Linguistic Treatment of Salience

The structure of marine weather forecasts shows several linguistic correlates of data salience relations. First, warnings of dangerous conditions (strong winds and freezing spray in the FPCN25 region) constitute separate headers preceding the normal text. Only warnings are so positionally marked and informationally redundant. Within the normal text, sentence groups dealing with each forecast parameter are ordered by two principles: intrinsic interest of the data and implicit causal links between the events or states described. Thus wind direction and speed, as the critical factors in marine conditions, occupy initial position. However visibility ratings, which should follow in order of importance, occur last by virtue of their dependence on fog/mist descriptions, which in turn are somewhat dependent on precipitation, which in turn follow cloud cover ratings. Sentence groups are therefore ordered as follows:

WINDS > CLOUD-COVER > PRECIP >
FOG&MIST > VISIBILITY

Within each sentence group, sentences and clauses are first ordered according to the dichotomy "general vs. local exception", and then chronologically within general and exceptional parts. A final correlate of data salience is the choice of marked lexical items and modifiers. For example, particularly strong winds are classified as "gales" (at 35 knots), "storm force winds" (at 45 knots), etc. Also, more specialized sense verbs such as "veering" and "backing" tend to be used more for large changes of wind direction.

7. Temporal Reference under Increasing Uncertainty

An interesting problem arises in ascribing particular time adverbials to points and intervals of (local) time. There appears to be a tendency in reports to "hedge" temporal descriptors slightly as reference time becomes more remote from the forecast issue time. For example, "Tuesday afternoon" or "by (Tuesday) evening" may be preferred for remote reference over the more precise "late Tuesday afternoon". This may reflect the increasing difficulty in predicting onset times for remote meteorological events. RAREAS incorporates two varieties of temporal rules in order to generate more vague temporal descriptors for more remote events.

8. Bilingual Reports

The RAREAS system was designed to accomodate the synthesis of marine weather bulletins in French as well as in English. Only the final three components in the processing sequence are language-dependent (and only the last of these in a non-trivial way). Syntactic patterns and lexical entries for French must of course be furnished on the basis of independent linguistic study of the corresponding French sublanguage. The exact semantics for French (correspondences between data configurations and specific lexemes) must be worked out separately, since there is no guarantee that English and French are lexically one-to-one, even in this narrow domain.

Canadian weather forecasts of all varieties are currently translated into French by the METEO system (Chevalier et al. 1978), developed at the Université de Montréal some ten years ago. Although METEO takes advantage of the relative closure and stereotyped style of forecasts, a certain percentage of forecast sentences fails analysis and hence translation. This is due not only to input errors due to typing and line noise, but also to slight irregularities in the usage of English grammar and lexicon on the part of forecasters, which have proved troublesome to foresee in a compact system.

The automatic synthesis of marine forecasts, on the other hand, should eliminate the fuzzy edges of unpredictability in human language production, by using a semantically complete and consistent subset of language to cover all foreseeable data configurations. Work on RAREAS thus prepares the ground for an attractive alternative to machine translation of these forecasts. The simultaneous synthesis of English and French forecasts directly from data would optimize the transfer of information to speakers of both languages, in addition to being (in principle) more reliable. Parallel synthesis of bilingual forecasts bypasses translation altogether, and most of the system's work in fact serves for both languages.

9. Implementation

RAREAS is written in MProlog, and runs on a Vax under VMS as well as on PC/XT/AT compatible microcomputers. Synthesis of a complete five-area forecast (about 150 words) takes about half a minute for the Vax implementation and a minute for the AT implementation. In either case, this is probably less than the time required to type or write the same forecast by hand, not to mention compose the same forecast from data.

10. References

Becker,J. (1975) "The Phrasal Lexicon", *Proc. TINLAP 1.* Cambridge, Mass.

Chevalier,M. et al. (1978) *TAUM-METEO.* Groupe TAUM, Université de Montréal.

Contant,C. (1986) "Génération automatique de texte: application au sous-langage boursier", M.A. thesis, Dépt. de Linguistique, Université de Montréal.

Harris,Z. (1968) *Mathematical Structures of Language.* Wiley-Interscience

Kittredge,R. and Lehrberger,J. eds. (1982) *Sublanguage: Studies of Language in Restricted Semantic Domains.* deGruyter

Kukich,K. (1983) "Design of a Knowledge-Based Report Generator", *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.*

McKeown,K. (1982) "Generating Natural Language Text in Response to Questions about Database Structure", Ph.D. thesis, University of Pennsylvania Computer and Information Science Department.