DEEP COMPREHENSION, GENERATION AND TRANSLATION OF WEATHER FORECASTS (WEATHRA)

by BENGT SIGURD, CAROLINE WILLNERS, MATS EEG-OLOFSSON
and CHRISTER JOHANSSON
Dept of Linguistics, University of Lund, Sweden
E-mail: linglund@gemini.ldc.lu.se FAX:46-(0)46 104210

Introduction and abstract

Weather forecasts were early noted to be a domain where automatic translation was possible (Kittredge, 1973). Everybody in the field knows that there is a computer in Montreal translating forecasts routinely between French and English (METEO). The weather domain has proven to be a fruitful domain for further research as witnessed e.g. by the system for generating marine forecasts presented by Kittredge et al (1986), by the work by Goldberg et al (1988), by the system generating public weather reports in Bulgarian reported on by Mitkov (1991) and the system translating Finnish marine forecasts into Swedish by Blåberg (1988).

The Swedish Weathra system to be presented in this paper explores the language and semantics of weather forecasts further and it aims at deep comprehension of forecasts. Beside grammatical representations, Weathra uses representations of the meteorological raw facts and secondary facts, e.g. the fact that it will probably rain at a place where there is a low pressure area. It uses a representation of meteorological objects with their properties as frames in a data base and graphic representation with the standard meteorological icons on a map, e.g. icons for sun, cloudy, rain, snow, thunderstorm, westerly winds, L(ow) and H(igh) pressure, temperatures, e.g. 10-15. Weathra also features a dynamic discourse representation including the discourse objects which may be referred to by the words and anaphora in the text (cf Karttunen, 1976, Johnson & Kay, 1990). The discourse objects are regarded as instances of the (proto)types or (concepts), which are also available as frames in a database.

The formal grammar, morphology and lexicon of Weathra are based on experience from the machine translation system Swetra (Sigurd & Gawronska, 1988), which is also written in Prolog (LPA MacProlog). The Weathra system can understand weather forecasts in a fairly deep sense, depict its comprehension in a map, answer questions about the main contents and consequences, translate English forecasts into Swedish ones and vice versa, and generate various forecast texts in English or Swedish.

The language of forecasts

Even a quick glance at the weather forecasts in newspapers shows that they are written in a special format beside using a restricted vocabulary (the METEO system uses some 1000 words, and so does Weathra). There are in fact two basic styles in the forecasts: the telegraphic style illustrated by *Sun; Cloudy; Windy; Cool; Morning fog; Thunderstorms in the coastal areas; High 20, Low 15; Westerly winds; Snow over Alps; Visibility moderate; Cloudy, little rain at first, brighter later* and a normal descriptive style illustrated by *A low pressure area is moving towards Scandinavia. It is expected to reach Norway in the afternoon.* There is, in fact, also an informal personalized style which may be illustrated by the following quotes from a British newspaper (the European):

Players in the Rugby League test match between Great Britain and Australia on Saturday may need longer studs and safe hands to tackle the tricky conditions at London's Wembley stadium. Sunseekers looking to top up their tan need look no further than southern Spain.

The Weathra system is primarily designed to treat the telegraphic and the descriptive styles.

The grammar of forecasts

Weathra includes two grammars motivated by the distinction between telegraphic weather phrases and full sentences. Interestingly enough the grammatical categories of the two kinds of expressions differ. The telegraphic phrase grammar can work with a superordinate category called *nominal* which includes both nouns and adjectives. The noun phrases used in telegraphic grammar may be somewhat different. They may, for instance, lack articles as evidenced by some of the examples mentioned above. The adjectives of the nominal category have a special marker *(-t)* in Swedish, cf English *sun in coastal areas*: *sunny in coastal areas* (Swedish: *Sol i kustområdet: Soligt i kustområdet.*

The telegraphic meteorological phrases lack finite verbs, but there is often a parallel full sentence with a future verb *will be* (available in the full sentence module).

1.a. Sunny in Wales
1.b. The weather will be sunny in...
2.a. Sun in the morning
2.b. There will be sun in...
3.a. High 20, Low 15
3.b. The temperature will be between 15 and 20
4.a. Visibility moderate
4.b. The visibility will be moderate
4.a. Probably rain
4.b. It will probably rain

The following is the basic rule showing how English weather utterances *(ewutt)* can be generated as phrases *(ewph)* or full sentences *(ewsent)*.

```
ewutt(T,F,S,[]) :- ewph(T,F,S,[]).
ewutt(T,F,S,[]) :- ewsent(T,F,S,[]).
```

The basic rule for phrases is:

```
ewph(A,[event(N),time(fut),advl(A),
      advl(B),co(C)]) -->
      eadv(A),enom(N),
      eadv(B),eco(C).
```

This rule generates e.g. *In the morning mild; Rain in the evening; In Scotland gale in the afternoon; Rain and snow,* i.e. nominal phrases with adverbial determiners before and/or after a nominal which can then be an adjective or a noun, single or coordinated. As can be seen the rules render a list representation called *functional event representation*, where the event is the first term, then actors, time and adverbials. There is normally no more than two adverbial phrases before or after the nominal. The last term *co(C)* takes care of cases of coordinated phrases or sentences. The first slot is used to indicate which constituent is in focus (first), information which is useful in the text generation process.

The further division of the superordinate category nominal *(enom)* is illustrated by the following rules for English:

```
enom(M) --> enp(Agr,M).
enom(M) --> eap(M).
enp(Agr,nom(H,Adj,Attr)) -->
eap(Adj),en(Agr,H),epattr(Agr,Attr).
```

Note that noun phrases have to carry agreement information *(Agr)* into the post attributive expression *(epattr)* as it might be a relative clause where the inflection of the verb depends on the features of the head of the np, as in: *light winds which turn west/ light wind which turns west.*

The following is one of the DCG rules generating and analyzing full sentences:

```
ewsent(N,[event(V),actor(N),
      time(T),advl(A1),
      co(C)]) --> enp(Agr,N),
      evi(Agr,m(V,T)),
      eadv(A1),esco(C).
```

It can for instance generate the sentence: *A low pressure area approaches Scandinavia.* Another pattern is used in order to generate sentences such as *The low pressure area will bring rain in Sweden,* etc. There are about a dozen different syntactic structures to be found in the forecasts.

The lexicon has the same format as Swetra. The first slot contains the form of the item (one or several words), the second the meaning written in "machinese", the third slot the grammatical category and further slots may be used for various features and classifications. The following are some examples.

slex([in],m(in],prep,_,_,_,_,_,_,loc).

slex(["Skandinavien"],m(scandinavia, prop),n,_,_,_,_,_,_,loc).

slex([på,eftermiddagen],m(in_the_af ternoon), adv,_,_,_,_,_,time,dur).

The lexicon includes a great number of multi-word-items of the kind illustrated by *på eftermiddagen*. These fixed phrases are particulary common in special domains such as forecasts.

The words, concepts and objects of forecasts

Consider the following text, with several alternative second sentences.

A gale is moving towards Scandinavia.

This sentence may be rendered in the following way in order to reveal the concepts and objects involved, some of which can also be referred to. The potential referential objects are numbered (within parentheses).

Something (O1), which is an instance of the concept 'gale' (O2) and is denoted by the English word gale (O3) does something (O4) which is an instance of the concept of 'move' (O5) and is denoted by the English word move (O6). The movement has a direction (O7) which is an instance of the concept 'towards' (O8) and is denoted by the English word towards (O9). The goal (O10) of the movement has the proper name Scandinavia (O11) in English.

The following are some possible successive sentences where the objects referred to are marked as O1,O2 etc.

It (O1) moves fast (1)
It (O4) happens fast (2)
It (O1) is better called a cyclone. (3)
It (02) translates as "storm" in Swedish (4)
It (O9) is better spelled toward (5)
It (010) includes Sweden (6)

We take a reference to prove that the object is a possible discourse object (*discourse referent* to use Karttunen's term, 1976). We may consider discourse objects as individual temporary mental objects created primarily for the sake of communication. They can be denoted by a word and are classified as instances of prototypes (concepts), when they are denoted by generic words. The classification is done according to a set of permanent prototypes (concepts). What the speaker does is typically to create temporary discourse objects, define them as instances of certain types (unless proper names can be used, hopefully known to the listener) and say something about them.

The first object created in our sample text is (O1) and it is said to be an instance of the type denoted by 'gale' in English. It is said to do something which is an instance of 'movement', etc. The second sentence may refer to different objects introduced in the first sentence, even to objects denoted by verbs or sentences *(It moves fast)*.

If we are to elaborate this sentence we would say *The movement (of the gale towards Scandinavia) happens fast*, but one may also use a hyperonym as in *The event happens fast* and possibly *The accident happens* (or *develops) fast*. We may say that the same object is being referred to in the sentence: *It is a disaster*, but in this case one may assume the existence of a type of object which has been called "inclusive referent" by Bonny Webber.

Note the distinction between a temporary object created in order to think or say something and a permanent object such as 'gale'. Even a concept can be referred to, as illustrated by the second successive alternative *(It translates as "storm" in Swedish)*, where *it* must refer to the object O2, which is a concept, not the object *O1*.

We may also note that a word used may be referred to, as illustrated by *It should better be spelled "forward"*, where *it* certainly refers to the word *forwards*. If *it* were to refer to the concept one would not use the word *spelled* but rather *denoted*.

This survey is intended to clarify that it is generally necessary to keep track of the following types of objects and representations:

1) Meteorological objects

These include both objects proper, such as low pressure areas and other air masses, and episodes (states, events, and processes) describing phenomena such as rain, change of temperature, as well as locations and time intervals.

2) Discourse objects

Discourse objects can be described by meteorological objects, but they also have linguistic expressions. Not all meteorological objects whose existence is implied by a forecast describe discourse objects.

3) Grammatical representations

Grammatical representations refer to expressions signifying discourse objects

The main levels of representation in Weathra are:

Level of meteorological objects

air mass: gale doing:move speed:
fast direction: Scandinavia

Level of discourse objects

O1:gale O2:move O3:Scandinavia
O4:(=O1) O5:move O6:fast

Level of functional event-repr.
[event(move),actor(gale),
time(pres),
advl(toward,Scandinavia)],
[event(move),actor(gale),time(pres),
advl(fast)]
Text level: *A gale is moving towards Scandinavia. It moves fast.*

The permanent concepts which constitute frames with information is only background objects alluded to by the words (cf Nirenburg & Defrise, 1991). The concept (frame) *gale* thus includes the information that a 'gale' has speed and that this speed is between 20 and 30 meters per second, a direction (which all speeds have), often leads to accidents at sea and along the coasts, etc. To an English-speaking person the concept is known to be denoted by the word *gale*, to a Swede by *storm*, but that is not essential information in the concept 'gale'. The concept 'move' is to include the information that movement implies being at one place first and another later, a certain speed and direction. To those who are familiar with it the concept 'Scandinavia' includes the information that this is a place and an area, which covers Norway, Sweden, Denmark etc. Scandinavia is a proper name and not a generic noun and something cannot be said to be an instance of the concept 'Scandinavia'. Concepts are stored as frames using the tool FLEX which is available with LPA MacProlog.

Understanding and generating weather forecasts

The program allows a (telegraphic or full) sentence to be parsed by the grammar and lexicon applying some implicational morphological procedures. This analysis renders a kind of functional representation as shown above. This representation is parsed by mapping procedures which look for depictable objects and places. Words such as *sun*, *sunny*, result in a *sun* in the proper place in a map, *rain* results in the proper symbol, *westerly winds* results in an arrow with the proper direction. Note that several words may result in the same symbol on the map. *Sunny, sun* and *fair* will all be represented by the icon "sun".

The functional representation is also scanned by the concept finder which looks for concepts about which it has information. Thus the frame 'gale' is used as the prototype of the instance O1 and 'move' is used as the prototype of 02. The meteorological finder looks for data for its general frames.

The system can be used for generation by placing a certain icon on the map and calling for generation. This will result in a sentence such as telegraphic: *Sunny* *in Southern Sweden, Sun in Southern Sweden, Fair in Southern Sweden*, or descriptive *It will be sunny.., There will be sun....The weather will be fair...*

The generation process may be set to generate telegraphic or full utterances, single or coordinated utterances, texts where the area is kept in focus, e.g. *Wales will get sun and light winds* or coordinations with different areas such as *Wales will get sun, but Cornwall will get rain.* The procedures may also generate texts where the focus is on the weather type as illustrated by *There will be snow in Scotland and in the Midlands* or *There will be snow in Scotland, but rain in Wales.*

Translation

The generation triggered by placing a meteorological icon on the map can be rendered in English or Swedish. Parsed and analyzed Swedish forecasts can be translated into English using the functional event representation. The functional representations of English and Swedish are very similar and the few differences are handled by transfer rules.

References

Blåberg, O. Translating Finnish weather forecasts into Swedish (Dept of Linguistics, Umeå: 1988)

Bourbeau, L, Carcagno, D, Goldberg, E, Kittredge, R & Polguère, A. Bilingual generation of weather forecasts in an operations environment. Proc. Coling 90, Helsinki (1990)

Goldberg, E., Kittredge, R & Polguère, A. Computer generation of marine weather forecast text. Journal of atmospheric and oceanic technology, vol 5, no 4, 472-483

Johnson, M & Kay, M. Semantic abstracting and anaphora. Coling 90. Helsinki (1990)

Karttunen, L. Discourse referents. In: McCawley, J (ed) Syntax and semantics 7, New York (Academic press:1976), 363-385

Kittredge, R, et al (1973) 'TAUM-73'. Montreal: Université de Montreal

Kittredge, R., Polguère, A. & Goldberg, E. Synthesizing weather forecasts from formatted data. Proc. of Coling 86, Bonn (1986)

Mitkov, R. Generating public weather reports. In: Yusoff, Z. Proceedings of the International conference on Current issues in Computational Linguistics, Penang Malaysia, 1991

Nirenburg, S. & Defrise, C. Aspects of text meaning. In: J. Pustejovsky (ed) Semantics and the lexicon. Dordrecht (1991:Klüwer)

Sigurd, B. & Gawronska, B. The potential of SWETRA - a multi-language MT-system. Computers and Translation 3, (1988), 238-250.