METEO, an operational system for the translation of public weather forecasts

Introduction

The TAUM project, from the University of Montreal, has been engaged for more than six years in the development of experimental models for the fully automatic translation of general texts from English into French. The first of these models has become known as TAUM 7I [I] and the latest one will be presented at the COLING 76 Congress. Because of the huge amount of data which needs to be compiled in order to make such a system, not to mention the introduction of a truly semantic component, no such system will be available for years to come. It is however possible to consider immediate limited applications for computer translation. METEO is an example derived from the TAUM 73 model. TAUM has also tried to demonstrate that computer translation could be successfully applied to the translation of technical manuals in a two-month experiment with the Canadian Translation Bureau and will concentrate in the next two years on the design of more appropriate parsing techniques and procedures for the treatment of idiomatic expressions.

General description

METEO is a fully automatic system for the translation of public weather forecasts from English into French covering the whole of Canada. It has been operating on an experimental basis since last December and due to be fully operational on the I5th of May I976.

Public forecasts for Canada are prepared in several regional centers from data sent by measuring stations throughout the country and centralized on a computer via the CN/CP communications network. Forecasts to be translated are retrieved, placed in a special file and processed one by one by the translation system. There is no human intervention prior to translation other than the actual typing of the text by a communicator at the corresponding regional office. The output of the translation system is handled by a specially designed editor (GERANT) which displays rejected sentences on a screen terminal at the local Translation Bureau. These sentences are taken care of by a human translator and as soon as a communication is completed it is redistributed to radio stations and newspapers using the same communications network as before. There is no

revision of the sentences accepted and translated by the system and to our knowledge this is the first time the product of a computer translation system will be distributed directly to the public. The sentences rejected by the sytem represent less than 20% of the total input, the main causes being: mispelled words, characters blurred in transmission, words not in dictionary, poor English, syntactic structures unknown to the parser, etc. The estimated load of the system is 30.000 words per day at the rate of over 1000 words per minute and the all-inclusive cost is about one third that of human translation.

The program

The program is divided in two main parts, the translation program and the editing program. The translation program is a succession of grammars of rewriting rules written in Q-System (2) Interpreters for this language are available in ALGOL, FORTRAN and COMPASS; the FORTRAN version was implemented on a CDC 7600 computer because it was judged to be the most transportable by both parties concerned. The editing program was written in FORTRAN for that particular application and also performs automatic preediting and formatting before and after translation.

The linguistic approach

The grammars are four in number:

The idiom dictionary
The main dictionary
The parser
The generator

The idiom dictionary

The idiom dictionary contains about 300 entries which can be divided into three types:

a) Several true idiom-like expressions such as:

clear period → eclaircie

b) A few strings of words which are not parsed for reasons of performance because they are compulsory elements of all communications:

"forecast issued by the atmospheric environment service"

c) A majority of place names which need to be translated or have an unpredictable translation:

Lake St Claire → lac Ste Claire

The main dictionary

the main dictionary contains all the lexical information necessary for parsing and generation and gives for each word the possible syntactical categories, for each category the possible translation or translations and for each category/translation pair the corresponding semantic features. Morphological variations of words also appear in the dictionary because there is only a very small number of them; in some cases the root form has even been omitted altogether, the infinitive of verbs for example. the present dictionary contains about I200 entries in all.

The parser

The originality of the METEO system lies mainly in the parsing techniques used. The world of weather forecasts is not unlike Winograd's blocks' world: lexicon, syntax and semantics are all restricted and make up a well-defined microworld. From the syntactic point of view, sentences are short and structurally simple, no relative clauses or passives for instance. The main problem is the delimitation of syntagms owing to the essentially telegraphic style of weather forecasts and the abundance of conjunctions. It was evident from the start that a conventional syntactic parser would be of little use because of the frequent omissions of function words and that it would be necessary to rely on some sort of semantic information. The ground work was laid out by. Richard Kittredge, Director of the TAUM project, in a preliminary study and a multiple-pass parser relying both on syntactical and semantic information was designed.

The aim of the parser is to give for each input string a single description giving the categories and translations realized in that particular string:

In a first pass substrings containing numerals are identified as dates, hours or temperatures.

In a second pass substrings expressing time or location are recognized as such. In the case of time, a distinction between durative and more punctual expressions is necessary because of the associated variations in French:

in the morning → dans la matinee

this morning → ce matin;

also, the distibution of determiners is often necessary when there is a conjunction:

this arternoon or evening →

cet apres-midi ou ce soir.

As far as locatives are concerned, the most difficult part is the identification of words not in dictionary as place names on the basis of context for place names which do not need to be translated were not entered in the dictionary because of their very high number.

In a third pass the remaining substrings are analyzed. The corresponding rules rely heavily on the semantic subcategorizations introduced in the dictionary to chose the proper translation for a given word:

heavy fog → brouillard généralisé

heavy rain → forte pluie

or to determine the scope of conjunctions:

snowflurries or rainshowers becoming intermittent tonight

(snowflurries or rainshowers) becoming ... .

For instance, it was necessary to divide weather conditions according to whether they were stationary, wind-like or precipitations in order to parse properly at this stage.

In a fourth pass the sequences of conditions, time references and locatives are tested for ambiguity and well-formedness and

each time a single structure can be built for a given input string, the parsing is retained and later processed by the generator.

In a fifth and final pass incomplete parsings are rejected and "styllistic" adjustments are made. An interesting example of this is the treatment of the word "occasional" which is entered in the dictionary as meaning "passager" because the predominant interpretation is the repetition in time, yet surely this is not the case for:

occasional cloudy periods -

*passages nuageux passagers

where one must assume that the meteorologist meant repetition in space, hence:

passages nuageux isolés

The generator

The task of the generator is to decompose the structure built by the parser, introducing articles where necessary, taking into account the word order of French:

gusty westerly winds →

vents d'ouest soufflant en rafales

and taking care of agreement.

Conclusion

The METEO system could not be used for the translation of texts other than meteorology because it is based on the semantics of that particular microworld but the strategy described here could certainly be adapted to limited fields where the amount of text to be translated largly compensates for the cost of designing a specific system. Neither do we wish to claim that our system is foolproof as demonstrated by a recent output:

apercu pour demain: faible possibilte

but then again, the communicator did type:

outlook for tomorrow: little chance.

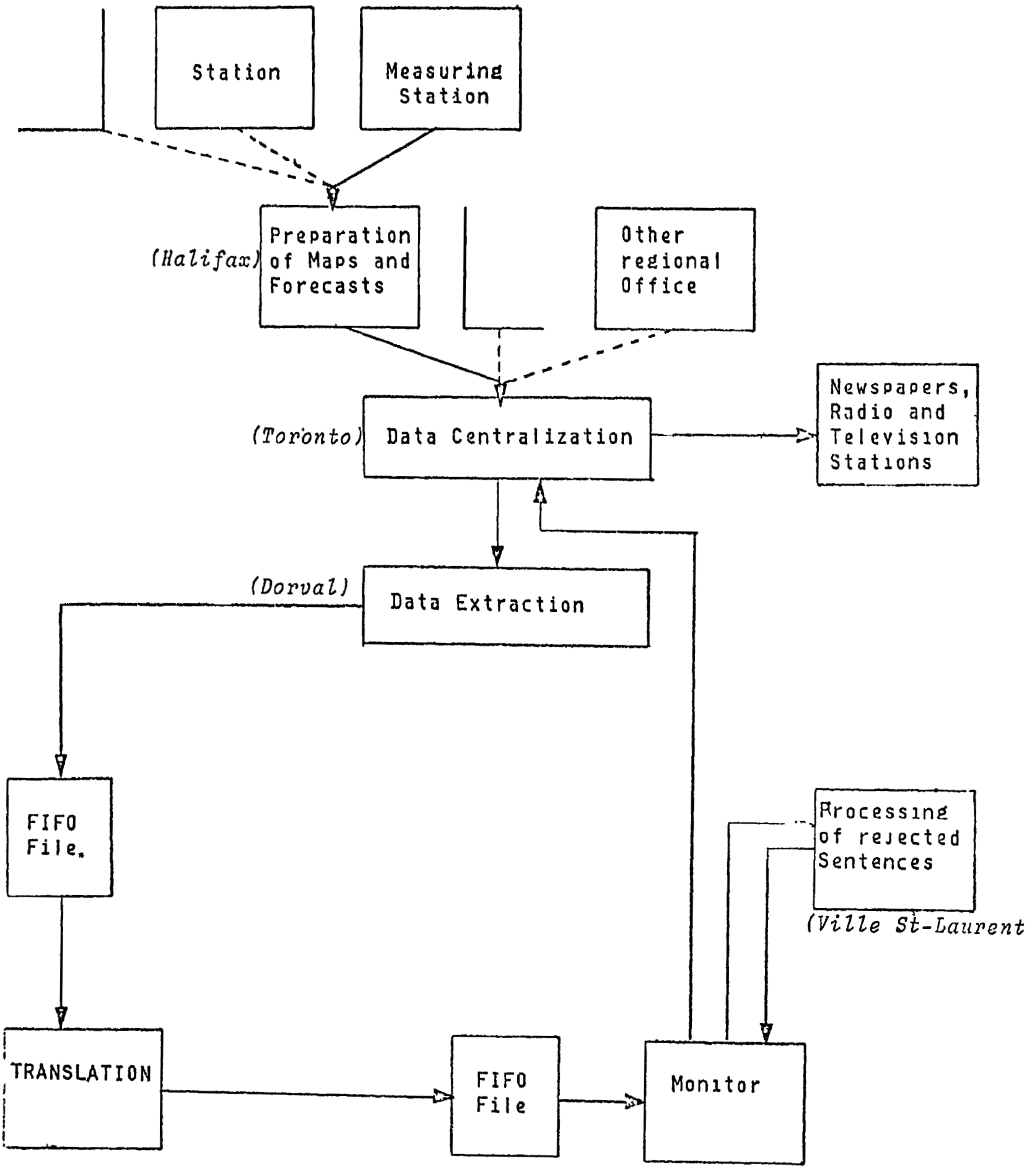
Nevertheless, we have found it to be a most entertaining project and our TAUM 76 model will benefit from it.

John CHANDIOUX
Head of the METEO team
TAUM project
University of Montreal

Chandioux received his Licence in English teaching in 1971 and in Linguistics in 1972, when he also received a Masters in English teaching; in 1973, he took a Masters in Linguistics. He is presently doing a Ph.D. in applied linguistics. Before joining TAUM he worked in France and Canada, teaching English as a second language and teaching contrastive linguistics.

(I) A. Colmerauer, Les Systemes-Q, université de montréal.

(2) TAUM 7I, université de Montréal.


Station
Measuring Station
(Halifax)
Preparation of Maps and Forecasts
Other regional Office
(Toronto)
Data Centralization
Newspapers, Radio and Television Stations
(Dorval)
Data Extraction
FIFO File.
Processing of rejected Sentences
(Ville St-Laurent
TRANSLATION
FIFO File
Monitor

HIGH LEVEL

WOOD BUFFALO REGIONS

MOSTLY CLEAR AND COLD WITH PERIODS OF VERY LIGHT SNOW TODAY AND WEDNESDAY. HIGHS NEAR MINUS 10 BOTH DAYS. LOWS TONIGHT MINUS 20 TO MINUS 22.

HIGH LEVEL

WOOD BUFFALO

AUJOURD HUI ET MERCREDI GENERALEMENT CLAIR ET FROID AVEC TRES FAIBLES CHUTES DE NEIGE PASSAGERES. MAXIMUM POUR LES DEUX JOURS ENVIRON MOINS 10. MINIMUM CE SOIR MOINS 20 A MOINS 22.

System Specifications

Languages Field Purpose	English to Fench Meteorology General Public
Pre-editing Post-editing Interactive editing	none none human translation of rejected sentences
GRAMMARS	Written in Q-Systems, a high-level programming language specifically designed for linguistic applications. Available in ALGOL, COMPASS, FORTRAN.
Dictionary	1200 rewriting rules loaded in central memory
Parser	300 rewriting rules bottom-up context sensitive
Generator (including Morphology)	300 rewriting rules
Maximum memory capacity required for loading and execution	60K words
Translation speed using the FORTRAN version on a CDC 7600 computer	1000 words per minute

Estimated operating cost (including human correction)	3.5 cents per word
Failures (including transmission errors, spelling mistakes and poor English)	Less than 20% of the input sentences.
Load	30,000 words per day
Operation	Has been operating 24 hours a day for 3 months on an experimental basis.
Delivery	May 76