[MT Summit III, Washington, DC, USA, 1-4 July 1991]

Applying an experimental MT system to a realistic problem

Pierrette Bouillon and Katharina Boesefeldt
ISSCO, 54 Route des Acacias
1227 Geneva, Switzerland
pb@divsun.unige.ch
kathy@divsun.unige.ch

Abstract

This presentation outlines the implementation of a machine translation system for avalanche warning bulletins in natural language, using a unification-based formalism developed at ISSCO, which will be introduced at the same occasion. Concrete examples taken from this project exemplify a modern approach to machine translation: a rich representation of the semantic content of a sentence, the use of a single grammar for parsing and generating as well as generation and transfer based exclusively on the semantic representation of a sentence. Simultaneously, the limits of bidirectional transfer are being tested.

1 Introduction

The aim of this presentation is to show how ELU,[1] a unification environment developed for linguistic experimentation with parsing, generating and translation, is being used for the elaboration of a concrete machine translation system dealing with a limited semantic domain.

In this paper we concentrate on the notion of a for mal representation as the basis for mapping between languages. We will show how a semantic representation consisting of a complex feature structure enables an efficient transfer between languages in this limited domain.

2 Translating avalanche bulletins

2.1 The project

The goal of the project is to implement a machine translation system capable of translating Swiss avalanche warning bulletins. These are prepared in German by the IFENA (Swiss Federal Institute for the Study of Snow and Avalanches)[2] a number of times a week during the winter season, the exact frequency of their appearance depending on weather conditions. They are then translated into the other official Swiss languages, French ant Italian. This research project aims to provide a more rapid, coherent and systematic translation of these bulletins.

2.2 The sublanguage of avalanche bulletins

The bulletins (cf, appendix 1) are an ideal application for a machine translation system; they deal with the limited domain of avalanches in Switzerland. This domain is well known and already partially formalized in the practical guide to avalanches [Salm, l982], where the general meteorological conditions and the different geographic regions of Switzerland are defined. It is thus possible to identify the special language used in the bulletins and to limit the system to the linguistic phenomena encountered in this sublanguage.[3]

As a result lexical, semantic and syntactic ambiguities can be considerably reduced and treated more easily. In a sublanguage the number of words and their meanings are limited and the same expressions are used repeatedly. Moreover the input is controlled in a very strict way as there is only a limited number of syntactic structures and it is neither necessary nor even desirable to express the same idea in different ways [Lehrberger, 1982].

The bulletins inform the population about the snow condition and the danger of avalanches in the different regions of Switzerland. They are generally divided into three subparts: a description of the general snow conditions, an estimate of the degree of danger in the different regions of the Swiss Alps and instructions about measures to be taken in these areas. In each of these parts, sentence structure and vocabulary are restricted, and many fixed expressions are used. Regarding this property of the bulletins, it. is thus possible to define semantic types allowing us to verify the coherence of all the phrases and possibly their role in a particular sentence as well as in the bulletin as a whole. The noun phrase *un grand danger général* ('a great overall danger'), for instance, is said to be correct because the noun *danger* and the two adjectives *grand* and *général* have been given the same semantic type, **danger**. It is also possible to identify semantic units which are the basic concepts of the world of avalanches. For instance, the French noun phrase *couverture de neige* ('snow cover') forms a unique concept and has thus to be considered a semantic unit. It is therefore stated in the lexicon that the lexeme *couverture* requires an argument whose value is *neige*.

[1] "Environnement Linguistique d'Unification"

[2] The project is partially supported by the IFENA.

[3] This formalization is done in collaboration with the avalanche experts of the IFENA.

The avalanche bulletins also present lexical and syntactic peculiarities. The words occurring in the bulletins (only about 1000) cover a restricted number of meanings and are used exclusively in certain structures. The syntax however, even though limited, is quite complex. In contrast to the relatively simple meteorological bulletins of the TAUM-METEO project [Lehrberger, 1982], the avalanche bulletins are written in running prose in the third person, and include simple and subordinate active, passive, positive and negative sentences in the present, past and future tenses, which can also be coordinated. Phrases of all categories can be very complex, including from a syntactic point of view well known attachment ambiguities, and also exhibiting rich possibilities of coordination.

-

3 The ELU MT development environment

ELU, an enhanced PATR-II [Shieber, 1980] style environment. developed at ISSCO, is a program which interprets grammars written in a special language [Estival, 1990, Russell et al., 1991]. As its name indicates, ELU is based on unification and uses a system of features and values, called feature structures, as its domain of information and as a means of representing grammatical and semantic properties of a sentence. The choice of a unification-based formalism brings the benefits of declarativity of grammars, and the possibility of applying formal methods developed in computational linguistics to the study of certain aspects of translation.

It is composed of three main modules: a parser, a generator and a transfer module. For a given sentence the parser creates some number of representations whose form is defined by the grammar. Starting from a representation of the sort created by parsing, the generator will attempt to apply the rules of its current grammar so as to discover the string(s) of words which the grammar relates to that representation. The transfer module establishes a binary relation over a set of feature structures, associating the analysis of one structure with the synthesis of another and thus permitting us to pass from a meaning representation designed with one language in mind to one suitable for another language. For instance, the following lexical transfer rule establishes a direct relation between part of the representation of the meaning of the words *Gefahr* and *danger:*

```
    :TA:   gefahr danger
```

These atoms appear as the values of the attribute **<head sem pred>**, i.e. the value of the predicate attribute in the structure containing **sem**antic information within the **head** component of the overall representation. In conjunction with the transfer rule '**pred**', shown below, this rule establishes a transfer correspondence between the German structure **<head sem pred> = gefahr** and the French structure **<head sem pred> = danger**. The rule '**pred**' can be interpreted as stating that, if the input structure contains the path **<head sem pred>** and the value of this path can be transferred by some rule, then the output structure will also contain a path **<head sem pred>**, whose value will be the result of transferring the value of the corresponding path in the input structure:

```
        :T: pred
        :L1:<* head sem pred> =  X1
        :L2:<* head sem pred> =  X2
        :X: X1 = X2
```

The translation process based on this transfer module consists of three steps: analysis of the sentence of the source language, transfer, and generation of the target language sentence. During the transfer process, feature structures resulting from the analysis of the source sentence are transformed by the transfer rules in order to obtain a correct representation in the target language. Starting from this representation the generator generates the sentence in the target language. Applying ELU to the concrete project of an avalanche machine translation system provides an opportunity to test a modern approach in machine translation, based exclusively on the semantic representation.

The possibility of using the same grammar for parsing and generating offers considerable advantages, even though this approach requires a certain amount of care [Russell et al., 1990]. In the first place, it is no longer necessary to write, test, and maintain two different but compatible grammars. Secondly, it provides the grammar writer with an ideal method for testing that the grammar does not accept illicit sentences. And, finally, it allows the grammar writer to ensure the coherence of the system: all sentences that can be parsed can also be generated.

The ELU system also permits the transfer relation to be made reversible; that is to say, if a set of transfer rules applied to a structure *A* produce (perhaps among others) a structure *B,* then applying the same transfer rules in reverse to *B* will produce (perhaps among others) *A,* Reversibility of transfer has the advantages mentioned above in connection with parsing and generation - economy of effort, and greater control in the development of descriptions.

In the project we are presenting, generation and transfer are exclusively based on semantic representations, Syntactic and morphological information which is not relevant for transfer is not taken into consideration. The semantic representation of a sentence only contains information about the predicate of the sentence, its argument(s), the tense (present, past or future), whether it is a positive or a negative sentence and which are its modifiers (cf. example 1, section 4). Representations are complex; those of the subparts of a sentence are embedded inside that of the sentence as a whole.

3.1 Current state of the project

After an exhaustive study of the corpus with the aid of IFENA, we are now focusing on the translation. Grammars for French and German are being written, which treat all the basic phenomena encountered in the avalanche corpus. They make it possible to parse and generate a large number of sentences of the corpus with different complexity (appendix II). At the same time we are writing the transfer rules which at present permit the bidirectional translation of sentences of the type shown in appendix III. Section 4 will discuss the formal representatimi of some of the sentences cited in the appendixes II and III. With respect to the examples given here, it should he pointed out that a small amount of preprocessing has been assumed: limitations of the current operating system (though not of the ELU system itself) prevent the input of accented characters, and these are accordingly replaced with a postscript integer (e.g. *à* becomes **al**, and *übrig* becomes **u5brig**); in addition, certain fixed expressions arc coalesced (e.g. *au dessus* beroines **au_dessus**).

4 Meaning representations

As the world of avalanches forms a sublanguage with very few structural variations between the different languages, we tried, whenever possible, to represent the German sentences and their translations with the same semantic structures. We therefore had to complicate the grammars for parsing, but, in doing so, we considerably simplified the transfer rules, thus gaining in overall efficiency. The transfer phase is in many ways the most complex, and it is preferable to minimize the amount of work involved.

The following are examples of semantic representations used for transfer. They show the expressive power of the feature structures representing the semantic contents of the sentences. The dense representation is due to the list structures (**[A,B]**) and the tree structures (**X(Y)**). Certain substructures are labelled for ease of identification (<*n*>); re-entrant structures are indicated by a value of the form **'=> #n'** (e.g. **#2** in example 2). Re-entrancy expresses the type of identity that occurs in representations of control constructions and coordinate structures. To obtain the semantic representation of the French text, atomic values have to be substituted according to the atomic transfer rules.

Example 1
(active sentence with altitude modifier and compound subject):

> *In der Westschweiz besteht oberhalb 2000m eine massige lokale Schneebrettgefahr*

translated into French as:

> *Un danger local modéré de plaques de neige subsiste au dessus de 2000m en Suisse Romande*

('A moderate local danger of snow patches exits above 2000m in the western part of Switzerland'). The following part of the feature structure specifies that the input has been analysed as a positive active sentence in the present tense with the verb *bestehen* as its main predicate and two prepositional phrases - altitude (**alt**) and place (**loc**) - as its modifiers. The empty lists '**[]**' indicate that no corresponding item (e.g. time modifier - the value of **temps**) has been identified during the analysis.

```
mod:alt:head:sem:pred  = oberhalb(m(2000))
    loc:head:sem:compl = []
                 detype = definite
                 mod:n_comp = west
                          nom  =   []
                         pred  =  schweiz
     temps  =   []
morph: temps = present
      voix = actif
positif  =  yes
pred = bestehen
```

The argument of the predicate is the noun phrase *eine mässige lokale Schneebrettgefahr.* In order to obtain the same representation in French and German and to generalize the transfer, German compound nouns have been given lexical entries that assign them a complex representation. The compound *Schneebrettgefahr* has therefore been assigned *Gefahr* as its head and the compound *Schneebrett* as its complement, itself consisting of the head *Brett* and the complement *Schnee.* Adjectival modifiers of this compound are encoded as modifiers of the main predicate **gefahr**. The information **detype = non** (i.e. 'no determiner') has been added to enable the generation of the French nominal phrase.

```
compl:head:sem:compl:head:sem:compl = []
                            pred  =  schnee
           detype= non
           pred =  brett
detype = indefinite
mod:n_comp:head:sem:pred = []
nom =  [ <1>
            head:sem:mod:degre = []
            pred = ma5ssig
            <2>
            head:sem:mod:degre = []
                     pred = lokal ]
pred = gefahr
```

Example 2
(coordinated predicative adjectives):

> *In den Alpen ist die Gefahr noch gering und lokal*

translated as:

> *Le danger est encore local et modéré dans les Alpes*

('In the Alps, the danger is still local and moderate'). In this example the two coordinated adjectives *gering und lokal* share the same argument *Gefahr* which is shown by means of re-entrancy - i.e. the structure labelled with **'<2>'** is the value of two distinct paths:

```
head:sem:args = [ <1>
                    head:sem:<2>
                    compl =  []
                    detype = definite
                    pred = gefahr ]
          mod:degre:head:sem:pred =  []
         pred = gering
 head:sem:args  =    [ <3>
                   head:sem => #2 ]
          mod:degre:head:sem:pred =  []
          pred =   lokal
```

The verb *sein* ('to be'), which in itself has no semantic value, has not been assigned a place in the representation. The predicate of the sentence is the operator **und** which has as its arguments the meanings of the two adjectives *gering* and *lokal.* There are place and sentential modifiers **alpen** and **noch**, but no height modifier.

```
mod:alt:head:sem:pred =   []
    loc:head:sem:compl =  []
                detype = definite
                nom =   []
                pred = alpen
     phrase  =  noch
     temps =   []
morph:temps = present
      voix = actif
positif  =  yes
pred = und
```

5 Conclusion

ELU represents an original, modern approach in machine translation because of the possibility of using the same grammar for parsing and generation, the bidirectional transfer mechanism, and generation and a transfer components based on representations that can be specially created for this purpose by the grammar writer. Feature structures provide a flexible, elegant and efficient method of expressing the morphological, semantic and syntactic information relevant for translation in a unified and condensed manner.

The goal of this presentation has been to show a concrete application of the ELU (Environment Linguistique d'Unification) software. After a brief description of the project, we have focused on the importance of the notion of a sublanguage for the development of a machine translation system. Limiting the domain makes it possible for us to define a coherent semantic representation of the sentence for a efficient transfer between German and French. Simultaneously, we have pointed out that our machine translation system benefits from the advantages of a modern unification-based approach.

References

[Estival, 1990] Estival, D. (1990) "ELU User Manual." Technical Report 1, ISSCO, Geneva,

[Estival et al., 1990] Estival, D., A. Ballim, G. Russell and S. Warwick (1990) "A Syntax and Semantics for feature-Structure Transfer," *Proceedings of The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language,* Austin, Texas, 11th-13th June.

[Lehrberger, 1982] Lehrberger, J. (1982) "Automatic Translation and the Concept of Sublanguage," in R. Kittredge and J. Lehrberger (eds.) *Sublanguage: Studies of Language in Restricted Semantic Domains.* Berlin and New York: de Gruyter. pp. 81-106.

[Russell et al., 1990] Russell G., J. Carroll and S. Warwick (1990) "Asymmetry in Parsing and Generating with Unification Grammars: Case Studies from ELU,2 *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics,* Pittsburgh, 6th-9th June. pp. 205-211.

[Russell et al., 199l] Russell, G., A. Ballim, D. Estival and S. Warwick-Armstrong (1991) "A Language for the Statement of Binary Relations over Feature Structures," *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics,* Berlin, 9th-11th April, pp. 287-292.

[Salm, 1982] Salm, B. (1982) *Lawinenkunde für den Praktiker.* Bern: Verlag des SAC.

[Shieber, 1986] Shieber, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar.* CSLI Lecture Notes No.4, Stanford University.

Acknowledgement

This article and the work on the project as a whole have benefited greatly from the detailed comments and suggestions of Graham Russell and Susan Warwick-Armstrong.

Appendix 1: Example bulletin (9th March, 1984)

Starke Sonneneinstrahlung und eine Erwärmung zu Beginn der Woche sowie die nachfolgende kräftige Abkühlung haben die Lawinengefahr weiter gemindert. Die Niederschläge der letzten zwei Tage brachten lediglich dem Alpennordhang 10 bis 30 cm Schnee.

Am Alpennordhang, im Wallis, in den Tessiner Bergen, in Nord- und Mittelbünden sowie im Unterengadin ist die Lawinengefahr dank einer stabilen Altschneedecke gering. Einzelne Gefahrenstellen können noch an Schattenhängen oberhalb rund 2000 m angetroffen werden.

Im Oberengadin und in den südlichen Bündner Tälern, wo das Schneedeckenfundament immer noch eine ungenügende Festigkeit aufweist, besteht für den Skifahrer vor allem an Nord- und Osthängen oberhalb rund 2000 m eine mässige örtlich beschränkte Lawinengefahr.

Appendix II: Sentences parsed and generated by the French grammar

- 20 à 30 cm de neige fraîche sont tombés dans les Alpes, dans la région du Simplon, dans les vallées de la Viège et dans la région de la Maloja. (transitive sentence in the past tense)
- Le rayonnement intense et l'élévation modérée de la température ont influencé le tassement de la couverture de neige. (transitive sentence with a coordinated subject)
- Les journées ensoleillées, mais froides ont influencé la stabilité de la couverture de neige. (coordinated adjectives)
- Les chutes de neige qui continuent, causent une nouvelle aggravation du danger d'avalanches. (relative clause)
- Une constitution défavorable de la couverture de neige incite à une prudence accrue sur les pentes raides.

- Le danger de plaques de neige est local et modéré dans les autres régions.
- Le touriste doit veiller à un danger élevé d'avalanches sur les pentes raides et ensoleillées où la couverture de neige est ramollie. (subject control)
- Il faut considérer que la neige fraîche empêche le refroidissement de l'ancienne couverture humide.
- Bien que les chutes de neige soient insignifiantes, le danger local de plaques de neige augmente. (subjunctive adverbial clause)
- Après que des avalanches sont descendues et qu'un refroidissement marqué s'est produit, la situation s'est calmée.
- Un danger local modéré de plaques de neige subsiste au-dessus d'environ 2000 m sur le versant nord des Alpes, dans la région du Gotthard, dans les Grisons et en Basse Engadine. (coordinated place modifiers)
- Les voies de communication sont menacées par des glissements de neige dans les bassins d'exposition sud. (passive sentence)
- Les endroits dangereux se trouvent sur les pentes raides et sur les cuvettes d'exposition nord-ouest à nord-est.
- Les hauteurs de neige qui sont très inférieures à la moyenne atteignent environ 10 à 50 cm sur des pentes raides de exposition nord à sud-est au dessus de 1000 m.
- Après le temps variable des derniers jours, les hauteurs de neige atteignent environ 30 à 50 cm à 1000 m sur le versant nord des Alpes, en Engadine et en Valais après une semaine pauvre en précipitations, (multiple temporal modification)

Appendix III: Sentences translated from German to French and from French to German

- In den Alpen, im Simplongebiet, in den Vispertälern und im Malojagebiet fielen 20 bis 30 cm Neuschnee. (coordinated place modifier)
- In den übrigen Gebieten beträgt der Neuschnee weniger als 10 cm.
- Am ganzen Alpennordhang, im Wallis, im Gotthardgebiet und im Unterengadin besteht eine sehr grosse allgemeine Lawinengefahr. (intransitive verb and complex nominal phrase)
- In den Alpen ist die Lawinengefahr noch gering und lokal. (predicative phrase and coordinated adjectives)
- Der kräftige Temperaturanstieg und die nachfolgende Abkühlung beeinflussen die Schneedecke.
- Am Alpennordhang und im Wallis betragen die Schneehöhen rund 20 bis 50 cm.
- Ein ungünstiger Schneedeckenaufbau ermahnt zu erhöhter Vorsicht.
- Am Alpennordhang, im Wallis, im Gotthardgebiel und im Engadin bestellt oberhalb 2000 m eine mässige lokale Schneebrettgefahr. (height modifier)
- In den übrigen Gebieten herrscht oberhalb 1800 m eine mässige lokale Schneebrettgefahr.
- Die grosseren angekündigten Schneefälle führen in Alpenkammnähe zu einem weiteren Ansteigen der Lawinengefahr.
- Die niederschlagsfreie Witterung und die hohen Temperaturen führten zu einer günstigen Setzung und Verfestigung der teilweise geringen Schneedecke.