Harakka T. & M. Koskela (toim.) 1996. Kieli ja tietokone. AFinLAn vuosikirja 1996. Suomen soveltavan kielitieteen yhdistyksen (AFinLA) julkaisuja no. 54. Jyväskylä. s. 149 – 155.

AUTOMATIC TRANSLATION AND TEXT GENERATION

Bengt Sigurd

Lund University

If you search for Machine Translation (MT) on Internet you will get some idea of the situation in the field. You will find offers by firms which say that they translate atomatically using the latest techniques, you will find advertisements for Translators Workbench a tool for human translators, you will find systems which use neural nets when translating and you will find translation projects which are mainly linguistic tests of the potentials of formal grammars. When prices are given there is generally a difference between completely automatic translation and more appropriate renderings, improved by post-editing. According to one offer automatic translation between English, French, German and Spanish is carried out for 3 cents a word, but if the text is to be "touched up by human editors" the price is 10 cents a word. The magazine *Byte* had a special issue January 1993 dealing with Machine Translation.

GRAMMATICAL AND LEXICAL APPROACHES

Since projects in machine translation were started in the 1950s there has been two main roads to follow: the grammatical road based on computerized grammar rules and the lexical road based on extensive lexicons including long phrases. The grammatical road made great progress thanks to the formal generative grammar introduced by Noam Chomsky in the 1960s. Chomsky was himself not at all interested in automatic translation, but his formalism made it possible to write grammars which computers could apply both in analysis and synthesis (generation). The modern programming language Prolog offers ways to write grammar easily and modern grammatical research is often done directly on the computer where the rules can be tested immediately. The grammatical approach, however, mainly satisfies the linguists, as they take an interest in ways of handling such matters as self-embedded relative clauses, complex verb phrases with several auxiliaries and pronominalization. But the consumers are often less impressed as what is

linguistically interesting is often very rare in empirical texts and to the consumer it is the quality of the real texts that counts.

The lexical approach stems out of the experience that it is mainly the lexicons that are used by human translators and what human translators do is very much trying to find the proper words and phrases. The lexical approach can be extended to a phrase-to-phrase approach and even a sentence-to-sentence approach. In the modern version called "example based automatic translation" the computer tries to translate whole sentences and even longer sections first looking for a previous example of its translation in the data base. Next the computer looks for shorter examples of translated parts eventually ending in word-to-word translation. Many automatic translation systems have a word-for-word mode to which they resort when the syntactic analysis has failed. Generally they also have a no-translation escape, which means that the word is rendered just as in the source text. Many modern systems also utilize knowledge of the frequency of different constructions, phrases and words.

It is a fact that phrases and even sentences often recur in texts within a certain domain and if you look at parallel texts as e.g. the texts in French and English from the Canadian parliament you will be able to find a great number of phrases, sentences and even sections which you can reuse in translation. Finding and marking the equivalent parts of parallel texts is called alignment and there are some current projects working in this area.

The new fast computers with enormous memory capacity have made the example-based approach more feasible. In the example-based approach there is no grammatical analysis and no information about the categories, the subject, the head of the noun phrase, etc and word order and agreement may easily come out incorrectly. Consumers must, however, balance these errors against the errors found also in the sophisticated translation based on detailed grammatical and semantic analysis. For restricted domains such as weather and stockmarket bulletins example-based translation offers quick (and dirty) translation often with few and inconsiderable errors. An experimental example-based system (a phrase translator) has been developed at the Department of Linguistics, Lund University.

EUROTRA AND OFFSPRINGS

The EU project in machine translation EUROTRA is very well known. This large scale attempt to translate automatically between all the countries of EU failed, but a considerable know-how was gathered and there are off-springs in several places. In Denmark a center for language technology was established and there are several similar centers or research teams in Europe.

One system based on the Eurotra project is called CAT2. It was presented by Randall Sharp in an abstract in the following way at a conference in computational linguistics in Lisboa in 1994.

"CAT2
The CAT2 Machine Translation System, developed in Saarbrücken in 1987 is a natural language application coded entirely in Prolog. Developed initially out of the Eurotra Project, CAT2 has progressed to the point of being used in preindustrial and academic applications, using real-life texts and approaching large-scale linguistic coverage. CAT2 makes use of a simple formalism for expressing well-formed structures at any number of levels of representation, and for expressing the transformation of structures between levels.....Translations of structures which cannot be performed using the existing grammars and transformations are performed in robust mode, in which the substructures making up the structure are independently translated: this results in CAT2 always producing a translation of any input text, and allows the linguist/translator to concentrate on developing and refining specific linguistic phenomena. A specific application at the Universidad Nacional Autónoma de Mexico is briefly described, in which documents retrieved over the Internet via the Gopher server can be fully automatically translated by CAT2 on request."

The following passage from a news bulletin translated from Spanish into English indicates the quality of the translations by CAT2.

> TEHERAN, 1 de marzo (AFP)
> Dos personas murieron y mas de 20 resultaron heridas en un fuerte sismo registrado el martes de madrugada en varias localidades de la provincia de Fars, en el sur de Iran, anuncio Radio Teheran.
>
> El terremoto, de una intensidad de 5.7 grados en la escala de Richter, dano en un 50 por ciento una decena de pueblos entre Firuzabad y Farach-Band, 80 kilometros al sur de Chiraz, capital de la provincia de Fars, segun la radio.
>
> TEHERAN, 1 of March (AFP)
> Two persons died and more of 20 resulted heridas in a strong earthquake registered the Thursday of madrugada in various localities of the province of Fars, in the south of Iran, announced Radio Teheran.
>
> The earthquake, of an intensity of 5.7 degrees in the scale of Richter, damage in a 50 percent a decena of people among Firuzabad and Farach-Band, 80 kilometers at south of Chiraz, capital of the province of Fars, according to the radio.

'The translation of this section took about 1 minute. Since we are not familiar with the details of the system we will not comment on or try to explain the result.

SWETRA

The Swetra project at the Department of Linguistics, Lund is based on a grammar (The Swetra Referent Grammar including lexicon and morphology) constructed in order to investigate the feasibility of automatic translation. It has mainly been applied to the restricted domains weather and stockmarket. Swetra is an interlingua system, where the contents of the source text is represented in a formal (universal) language from which it should be possible to generate equivalent texts in several languages.

The following is an example of its potential when translating from Swedish into English (from Sigurd, 1994).

> *Årets sista börsdag slutade nedåt under måttlig omsättning. Affärsvärldens generalindex slutade på 1499 en nedgång med 0.3 procent. För helåret blev uppgången hela 54 procent, medan börsomsättningen ökade med 94 procent. Torsdagens omsättning blev 1400 miljoner. Astra A förmådde inte avsluta året på sin toppnivå.*
>
> *The last trading day of the year closed downward during moderate trading. The Business World's general index closed at 1499 a decrease by 0.3 percent. For the whole year the increase was as much as 54 percent, while the trading increased by 94 percent. Thursday's trading was 1400 million. Astra A could not close the year at its top level.*

There are a number of recurring words and phrases in the stock market bulletins, which make them suitable as a testing ground for MT systems. But sometimes the stockbrokers give more informal comments which are quite difficult to translate.

TEXT GENERATION

Automatic translation consists of the analysis of the source text and the subsequent synthesis of the target text. The contents to be expressed are then given by the source text, but the text synthesis or generation part of such systems can often also be used to express contents which have not been encoded in a natural language before.

Text generation is a special field of computational linguistics and it has its own section at international conferences in computational linguistics and its

own workshops. Typical generation programs comment on the locations and movements of objects in a scene (scene despription) as can be exemplified by the system Commentator (Fornell, 1983) developed at the Department of Linguistics, Lund. In one version it commented on the movements of airplanes and ships and could produce a report as the following:

The fighter is to the southeast of the cruiser.

The passenger airliner, which flies towards the east, appears from the west.

The fighter still flies towards the northeast.

The cruiser still heads towards the southeast.

The fighter disappears towards the east.

Now the cruiser heads towards the east.

Such situations can be depicted on a map or a series of maps, but in certain situations verbal reports have advantages. The problems of text generation are generally summarized by the following questions:

What to say?

In what order to say it?

How to say it?

There is always a lot to say, but only some interesting and relevant items should be mentioned (following Grice's maxims). As is obvious even from the short text above, the reader may easily get confused or bored. The order problem is often described as a linearization problem and it concerns the order of concepts within clauses, sentences, paragraphs and longer texts.

The problem How to say it has a lexical component where the conditions for using words are specified in such a way that the computer can apply them. The specification of these condition requires careful investigations and descriptions much more complicated than those found in traditional dictionaries. In order to use the words *still flies* the program must know that the airplane has moved before, i.e. that it has had different positions before and in order to say that it *flies towards the east (eastwards)*, the computer must determine the direction as demonstrated by the change in the coordinates of previous positions.

The question How to say it? also includes determining whether an indefinite article should be used - only for new objects - and whether the definitie article or a pronoun should be used - sometimes a reflexive pronoun. Only definite noun phrses are illustrated in the example above.

The following is an example of a text (overview) which can be generated from the same interlingua representation both in Swedish and English by one of the offsprings of Swetra (cf. Sigurd, et al. 1992):

> *Ett mäktigt högtryck har sitt centrum över Svealand. Det förskjuts snabbt åt öster och förstärks. Väster om högtrycket blåser en måttlig till frisk, byig sydostlig vind under förmiddagen. Under dagen blir det nästan klart. Under morgonen får vi lokal dimma.*
>
> *A massive high pressure is centered over central Sweden. It drifts quickly eastwards and strengthens. West of the high pressure a moderate to fresh, gusty southeasterly wind is blowing during the morning. During the day there will be almost clear weather. During the morning we will get local fog.*

The planning of weather texts includes determining which synoptic objects (low or high pressures, cold or warm fronts, etc) which determine the weather should come first. Then the information about wind, clouds, precipitation and temperature in the different areas should be ordered. The input to Weathra is meteorological data and facts.

The following is an exampel of a text generated by a sub-project of Swetra called Stocktra. It takes data from the stockmarket tables as input, makes various calculations, relates the movements of the shares and the bonds and relates the situation in Stockholm to the situation in New York and European stock markets.

> *Stockholmsbörsen vände neråt på onsdagen och räntorna gick upp. De europeiska börserna föll också. Affärsvärldens generalindex blev 1583, en nedgång med 0.5 procent. Den svenska kronan var oförändrad. Vinnare idag blev Astra B. Astra B steg med 0.5 kronor, en uppgång med 0.2 procent. Förlorare blev Aga A. Aga A sjönk 1.5 kronor, en nedgång med 2.1 procent. Branchvinnare var läkemedel. Förlorande branch blev industri. 2 höjda köpkurser noterades, 5 sänkta och 1 oförändrade. Det totala börsvärdet blev praktiskt taget oförändrat 1816650000 kronor. Omsättningen var måttlig; totalt omsattes aktier för 2524214 kronor.*
>
> *The Stockholm stock exchange turned downwards on Wednesday, and the interest rates rose. The European stockmarkets fell too. The Businessworld General Index was 1583, a fall of 0.5 percent. The Swedish crown was unchanged. Today the winner was Astra B. Astra B rose by 0.5 crowns,*

an increase of 0.2 percent. The loser was Aga A. Aga A fell by 1.5 crowns, a fall of 2.1 percent. The best branch was medicals. The losing branch was industry. 2 higher rates were noted, 5 lower and 1 unchanged. The total value of the market was almost unchanged, 1816650000 crowns. The trade was moderate; in total shares for 2524214 crowns changed hands.

CONCLUSION

We will certainly see many new applications of computers and Information Technology in the future. We can be sure that we cannot envisage all of them just as we could not in the 1950s envisage the wide-spread use of computers in office automation, word processing and communication in the 1990s. I think there will be machine translation facilities offered routinely on Internet and I think some texts appearing in newpapers and magazines will soon be translated automatically by the news agencies. I also think that some texts that we will read or listen to in the future will be generated by computers. The use of synthetic speech and speech recognition systems will probably also be much more common.

REFERENCES

Johan Dahl. 1995. "Stocktext-J". Ex. paper. Dept of Linguistics, Lund University.

Bonny Dorr. 1993. *Machine Translation. A view from the lexicon.* Cambridge, Mass: MIT Press.

Jan Fornell. 1983. "Commentator". *Praktisk Lingvistik* 8, 1-63. Dept. of Linguistics, Lund University.

Bengt Sigurd (ed.) 1994. *Computerized Grammars for Analysis and Machine Translation.* Lund: Lund University Press.

" " 1995. "Automatic Generation of Stockmarket Reports". *Working Papers* 44, 1995, 145-157. Dept of Linguistics, Lund University.

Bengt Sigurd, Mats Eeg-Olofsson, Caroline Willners & Christer Johansson. 1992. "Automatic translation in specific domains: weather (Weathra) and stock market (Stocktra, Vectra)". *Praktisk Lingvistik* 15. Dept of Linguistics. Lund University.

Muriel Vasconcellos. 1993. "State of the Art: Machine Translation." *Byte,* January 1993.