[*Progress in Machine Translation*, ed. Sergei Nirenburg. Amsterdam: IOS Press, 1993.]

Multi-Lingual Machine Translation Systems in the Future

• HOZUMI TANAKA

1 Introduction

For the past 10 years, research activities in machine translation have been increasing in many countries, especially in Europe, Japan, U. S. A. and the Pacific Rim countries. The more machine translation systems we are going to develop, the more language pairs we have to deal with. The increase of the number of language pairs enables us to accumulate valuable experience for multilingual machine translation systems, which have been a dream of machine translation researchers for many years. In the following sections we would like to discuss what we should do in the next 5 years to develop better multilingual machine translation systems.

2 Design of InterLingua

In order to avoid redundant work in developing multilingual machine translation systems, the interlingua (IL) approach is to be preferred even though there have been a lot of controversy about this approach. Some researchers made objections to the approach because there was no concrete idea of how do construct an IL with which most researchers would agree. But we have to pay attention to the fact that some of them did not deny the possibility of IL but insisted that it wad too premature to develop multilingual machine translation systems based on the IL approach. They thought it would take a long time to get a consensus on IL and could not wait.

Researchers who took the strongest position, claimed that it was not clear whether IL which was common to any languages existed since concepts differ from culture to culture and from language to language. However, this theory cannot explain the fact that we can communicate our ideas and intentions through language translation. Understanding using translation seems to support the existence of common concepts which can be shared by any

person whatever language he uses. Such common concepts will be language independent and seem to form a core of the 1L lexicon.

Note that the common concepts are not always the primitives using which more complex concepts are described. In other words, the common concepts have to include both primitive and composite concepts. Schank (Schank 1975) claimed the existence of a set of primitive concepts, but such a set is very difficult to define in practice. We will come back to the problem soon.

It is important to discriminate between two types of concepts — language-independent concepts and language-dependent concepts. As the former concepts are common to all languages, it is possible to include them in the core of the lexicon of IL as mentioned before. Although, at first glance it seems to be incompatible with the scheme of IL, we have to include the latter type of concepts in the IL lexicon because language-dependent concepts reflect the difference among different cultures where different languages are used.

In summary, the lexicon of IL will be composed of:

1. Language-independent and language-dependent parts;
2. Primitive concepts and composite concepts.

The most simple concepts in IL lexicon will be word senses contained in monolingual dictionaries but we have to establish correspondence between word senses of different languages. Will it be carried out manually? The careful manual task will lead to the best results but will be time consuming. Although in the end we might have to rely on our abilities to understand language, there will be another method with which the correspondence between word senses can be extracted automatically by using existing translation dictionaries such as English to Japanese and Japanese to English dictionaries. We discussed the problem in (Tokunaga and Tanaka 1990). Also we have to pay an attention to the work done by Ravin (Ravin 1989).

The EDR project, which was born from the Japanese Fifth Generation Computer Project, has been acquiring a huge collection of concepts which EDR calls the EDR dictionary. After the publication of this dictionary, more progress in the field of multilingual machine translation systems will become possible (EDR 1990) and many researchers will either enjoy or suffer from handling a huge volume of concepts in this dictionary. But they will have a very good experience for developing better natural language processing systems including multilingual machine translation systems.

In addition to the IL lexicon, what we would like to discuss next is the definition of relations between concepts in the IL lexicon. As you know, any concept can not exist by itself. In other words, it will have some relation to other concepts (Roget 1982; Sowa 1984; Lenat et al. 1986). The problems have been discussed in knowledge representation area of artificial intelligence (AI) but they have not given us a concrete set of relations between concepts. Lacking such a kind of concrete set might not cause any problem theoretically but it might cause a serious problem if we want to develop a real translation system.

With respect to the meaning structures of natural language sentences, a set of deep case relations have been often used, which is one of the important relations among concepts. However, as many linguists pointed out, it is very difficult to define a set of deep cases (Somers 1987). One of the reasons is that linguists want to define a small set of deep cases with which larger set of linguistic phenomena will be able to explain. On the contrary, machine translation researchers tend to use a large set of deep cases.

From the theoretical point of view, it is better to define a small set of deep cases but one often encounters exceptions in real texts which are very difficult to explain without modifying

the existing theory to introduce new case relations. Accordingly, most machine translation researchers have adopted a comparatively large set of deep cases. It is not necessary for us to confine ourselves to a small set of deep cases which some linguists have proposed.

The CICC project has proposed a large set of deep cases in order to cover linguistic phenomena in Japanese, Malaysian, Indonesian, Thai and Chinese. Even though linguists will have complaints about this approach, we must follow the MT-oriented approach of CICC. It reminds me of Martin's criticism against a system based on a small set of primitive concepts such as Schank's system. Martin once advocated his knowledge-based system OWL which was based on a large knowledge base (Martin 1975). As the CICC's proposal (CICC 1990) includes not only a set of deep case relations but also other relations, it will be a good starting point to develop better IL specifications.

In addition to deep case relations, ISA relation is a very important relation in IL which forms a hierarchical concept structure. The EDR concept dictionary contains such a hierarchical system which will play very important role to develop the next generation natural language processing systems including multi-lingual machine translation systems.

3 Standardization of IL through International Cooperation

Needless to say, the extraction of IL lexicon and relations between concepts will be possible only through international cooperation such as the CICC project. It should be noticed that each entry in the IL lexicon is written in IL and thus IL should be fixed at an early stage.

However the experience of the CICC project revealed that international cooperation is very difficult without a common base. Specifically, the ISA conceptual hierarchy is very important as the base of discussions of IL. According to my intuition, the upper portion of the conceptual hierarchy seems to be shared by many languages, and hence at least this part should be standardized through international cooperation early on. It will be a first step to have a better standardized IL which will give us a great impact to develop better multilingual machine translation systems.

In order to make international cooperations more effective, it is necessary to develop multilingual text corpora, which are not only useful to make comparison to a set of concepts in different languages but also necessary to evaluate a multilingual machine translation system.

4 More on Analysis

Apart from multilingual machine translation systems, the most important problem we have to solve in the next 5 years is the analysis of input sentences. To get the IL out of the input sentences we have to analyze the input deeper but we still lack better theories and algorithms for semantic and context analysis. Some researchers insist that the deeper the analysis the less information we will obtain. But this is totally incorrect. Instead we would like to say, the deeper the analysis, the more information we will get.

In the past, AI researchers have dealt with rather narrower domains and were able to carry out deeper sentence analysis. MT researchers have been just the reverse of AI researchers. They have dealt with broader domains but carried out shallower analyses. What is needed for the future MT research and development is to realize a deeper sentence analysis for the broader domain in collaboration with AI researchers.

We would like to make a point on the context analysis especially on the anaphora resolution. In 1980s, promising theories such as DRT (discourse representation theory), SS (Situation Semantics) and FCS (file change semantics) were proposed, but none of them could solve many important problems like anaphora resolutions. For instance, in the case of DRT, it certainly solved the difficult problem of quantifiers in the 'donkey' sentences, but it will work only after finishing anaphora resolution which is just what we want to solve. DRT did not give us any answers about how to perform anaphora resolution.

In the case of English, Sidner and Carter (Carter 1987) have proposed an interesting theory of anaphora resolution but their theory is not applicable to the other languages such as Japanese. It might be necessary to have a language by language theory to solve this problem.

5 Conclusion

In this paper we can only discuss few important problems in developing better multilingual machine translation systems. We emphasize the importance of developing IL dictionaries and multilingual text corpora through international cooperation. We also point out the need for the standardization of IL. In a few years, we have to begin international cooperation to discuss such problems as conceptual hierarchies and deep case relations in IL. In order to do so, it might be better to organize an international institute. Finally we discussed the necessity of developing more methods for deep analysis of text, including such phenomena as anaphora resolution. All of these problems are difficult to solve but without coping with these problems, we will not be able to have better multilingual machine translation systems.

References

Carter, D. 1987. *Interpreting anaphora in natural language*. Ellis-Horwood.

CICC. 1990. A specification of inter-lingua.

EDR. 1990. Concept dictionary. TR-027, Japan Electronic Dictionary Research Institute, Ltd., April.

Lenat, D., M.Prakash and M.Shepherd. 1986. CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *The AI Magazine*, **6**(4):65-85.

Martin, W.A. 1975. Semantic cases and characteristics for arguments to activities. Owl note, MIT AI Lab.

Ravin, Y. 1989. Synonymy from a computational point of view. Research Report 14962, IBM.

Roget. 1982. Roget's thesaurus (New edition)

Schank, R.C. 1975. *Conceptual information processing*. North-Holland/Amsterdam Elsevier.

Somers, H. 1987. *Valency and case in computational linguistics*. Edinburgh University Press.

Sowa, J.F. 1984. *Conceptual structures*. Addison-Wesley.

Tokunaga, T. and H.Tanaka. 1990. The automatic extraction of conceptual items from bilingual dictionaries. In *Proceedings of the Pacific Rim International Conference on AI*, 304-309.