Semantic Language and Multi-Language MT Approach Based on SL

GAO QingShi (~), HU Yue (~), LI Li (~) and GAO XiaoYu (~)

1 University of Science and Technology, Beijing, Beijing 100083, P.R. China
2 Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, P.R. China

E-mail: qsgao@public.bta.net.cn

Received July 15, 2002; revised May 20, 2003.

Abstract Any natural language is regarded as a representation of semantic language. The translation between two languages \((I, J)\) is regarded as a transformation between two representations. A natural language-I is translated into another natural language-J only by two steps. One is semantic-analysis of language-I based on "semantic-element-representation-base of language-I", the other is deploying into the representation of language-J based on "semantic-element-representation-base of language-J". For translating in N natural languages, it is needed to develop N translation systems only, rather than \(N^2\) or 2N systems.

Keywords natural language, machine translation, semantic element, representation of semantic element, semantic analysis

1 Introduction

The history of machine translation is more than a half century long, but its quality is still too poor. The key lies in linguistics, that is, the knowledge of language cannot be expressed in precision. N. Chomsky [1,2] is a precursor who applies mathematics to linguistics. Although there are controversies about his theory of linguistics, it is universally acknowledged that his formal grammar is a set of very important tools.

People conversing in different languages can understand each other, and different languages can be translated into each other, because their sentences have the same meaning. Generally speaking, it is not sure that words and phrases in different languages can be mapped to each other. In 1980’s, ICT of CAS proposed a new concept, meaning element and language model[3], i.e., semantic element representation (SER) with variables and SER without variables. For example, (subject clause) (sentence subtracting subject), as (part of adverb) as possible. Here (subject clause), (sentence subtracting subject), (part of adverb) are all variables. The value of a variable is a representation of semantic element, i.e., a sentence consists of representations of semantic elements which intersect each other rather than being arranged linearly. It is obvious that not all semantic elements have their representation in all natural languages. For example, “叶公好龙” does not have its representations in English. (comment: “叶” (Ye) 公 (Mr.) 好 (like) 龙 (dragon)” is a Chinese idiom.)

The concept of chunk was proposed by Prof. G.A. Miller of CMU from the view of cognitive psychology. For example, a number with 12 digits is your close friend’s phone number and the other is a digit of one. The difference of amount of information is significant, but both are one chunk. In 1980’s, concept of chunk was applied to a problem solving system SORT. It was applied to natural language by H.A. Simon before 1983[3] and by S. Abney in 1991[4]. We can consider semantic element as a chunk of language knowledge.

Although sentence-pattern-based is an old approach, sentence-pattern-based and template-based[5] are both good references. Some graduate students at Institute of Computing Technology, the Chinese Academy of Sciences, began to develop an English-Chinese machine translation system based on English-900 sentence-pattern with variables in 1985, and Prof. Wang Wen-Cai, Beijing University of Sciences and Technology, led a group to develop a Chinese-English-Russian machine translation system based on Chinese-900 sentence-pattern with variables. These are two early simple examples of sentence-pattern-based and template-based approaches in China. We can also consider seman-

This work is supported by the National Natural Science Foundation of China (Grant Nos.60083008 and 60273016) and the Foundation of Institute of Computing Technology, the Chinese Academy of Sciences (Grant No.20016250).
tic element as a pattern or template with recursive substitutions.

Some people regard rule-based and corpus-based as two different approaches. In practice, rule comes from corpus. Essentiality of rule is simplification. A Comprehensive Grammar of English Language[6] was written from end of 1950’s, when a specific institution about Survey of English Usage was established in London University, to 1985. More than ten million corpus items were collected during 25 years. The Oxford English Dictionary[7] was compiled by 500 scholars from 1857 to 1928. A large corpus from 25,000 books was collected. Corpus has taken the place of individual collections. In fact, it is impossible for any good book about language grammar to include all language rules.

Statistical method[s] is an important auxiliary method, for example, in collating language knowledge, solution by retrieval, etc. But it is not an independent approach, because it is not correct sometimes.

2 Semantic Language

Semantic of a sentence is called SS. An element to express a meaning in an SS is called semantic element (SE). SS consists of SEs. The representation of an SE in a natural language-I, such as English, Chinese, ..., is called the representation of semantic element in Language-I (SERi). A sentence, called SSRi, in Language-I is the representation of an SS in Language-I. People conversing in different languages can understand each other, and different languages can be translated into each other, because all SSs can be represented in different natural languages. Natural language-I consists of all SERis, including all SSRis. Semantic Language (SL) consists of all SEs, including all SSs. A natural language can be regarded as a representation of SL.

The formal definition of Semantic Language is as follows.

Semantic Language: \{ Ao, #, Eo, C, \phi \}

Ao: list of terminal symbols.
#: parameter identifier.
Eo: SE-set. \( E_0 = S_0 \cup U_0, S_0 = \{ \alpha | \alpha \in (A_0)^* \} \), \( U_0 = \{ V(X_V) | V \in (A_0)^*, X_V := \#C_{iv} | X_V \}
\#C_{iv} := C_{iv} | C_{iv} / C_{iv}, C_{iv} \in C_v, C_v \subseteq C \} \).
C: set of type. Each SE corresponds to one and only one type.
\( \phi: \#C_{iv} \rightarrow \alpha, \) where \( T(\alpha) \in C_v \);
\( \phi: \#C_{iv} \rightarrow V(X_V), \) where \( T(V(X_V)) \in C_v \).

Here \((X)^*\) is a part of the words in the list \( X \), \( S_0 \) is a set of \( \alpha \), and \( \alpha \) denotes an SE without any parameter. \( U_0 \) is a set of \( V(X_V) \), \( V(X_V) \) denotes SE with parameters. \( X_V \) denotes the parameter list of \( V \), \( C_{iv} \) denotes type-list, \( \phi \) is a set of parameter variable’s replacement rules, \( T(x) \) is the type of an SE \( x \), \( X \rightarrow Y \) denotes \( X \) can be replaced by \( Y \). A parameter variable \( \#C_{iv} \) can be replaced by a semantic element \( \alpha \) or \( V(X_V) \) if the type of \( \alpha \) or \( V(X_V) \) is equal to one type in type-list \( C_v \).

An SS expression, to represent the SS, is an SE, for which all parameters are replaced by SE without any parameter.

The formal definition of a natural language is as follows.

Natural Language-I: \{Ai, #, Ri, C, \phi\}

Ai: list of terminal symbols.
#: nonterminal symbol identifier.
Ri: SER-set. \( R_i = S_i \cup U_i \cup V_i, S_i = \{ \alpha | \alpha \in (A_i)^* \}, U_i = \{ \#C_{iv} | C_{iv} := C_{iv} | C_{iv} / C_{iv}, C_{iv} \in C_v, C_v \subseteq C \}, V_i = \{ \gamma | \gamma \in (A_i \cup U_i)^* \} \).
C: set of type. Each SER \( X \) corresponds to one and only one type, called \( T(X) \).
\( \phi: \#C_{iv} \rightarrow \alpha, \) where \( T(\alpha) \in C_v \);
\( \phi: \#C_{iv} \rightarrow \beta, \) where \( \beta \in U_i, T(\beta) \in C_v \);
\( \phi: \#C_{iv} \rightarrow \gamma, \) where \( T(\gamma) \in C_v \).

Here \((X)^*\) is a part of the words in the list \( X \), \( \#C_{iv} \) is called nonterminal symbol, which consists of the nonterminal symbol identifier \( \# \) and a type-list \( C_v \). Type-list \( C_v \) consists of limited (disorder) types \( C_v \), \( \#C_{iv} \rightarrow \beta \) or \( \gamma \) if the type of \( \alpha \) or \( \beta \) or \( \gamma \) is equal to one type in type-list \( C_v \).

Comment 1. Replacement rules are not productions. \( S_i, U_i \) and \( V_i \) are not produced by productions too.

Comment 2. In Semantic Language, it is unnecessary to distinguish different SEs corresponding to \( U_i \) or \( V_i \).

A sentence in a natural language-I is an SERi, for which all parameters are replaced by SERi without any variable.