COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

AN ENGLISH JAPANESE MACHINE TRANSLATION SYSTEM OF THE TITLES OF SCIENTIFIC AND ENGINEERING PAPERS

Makoto Nagao, Jun-ichi Tsujii (Kyoto University)
Koji Yada (Electrotechnical Lab.)
Toshihiro Kakimoto (Fujitsu Co.)

JAPAN

The title sentences of scientific and engineering papers are analyzed by simple parsing strategies, and only eighteen fundamental sentential structures are obtained from ten thousand titles. Title sentences of physics and mathematics of some databases in English are translated into Japanese with their keywords, author names, journal names and so on by using these fundamental structures. The translation accuracy for the specific areas of physics and mathematics from INSPEC database was about 93%.

1. INTRODUCTION

There have been many researches on syntactic analysis of natural language by computer, but still no reliable grammatical rules are established yet which can be applicable to any utterances of a language. Universal grammatical rules for a language looks like almost hopeless. Grammatical rules to be prepared depend heavily on the text to be analyzed. Hence the concept of subgrammar is introduced. It does not necessarily cover all the different kinds of sentential structures of a language. A grammar which covers just the set of expressions to be treated is sufficient from the engineering point of view.

We developed a machine translation system which translates the titles of scientific and engineering papers from English into Japanese. More than 98% of the titles in scientific and engineering papers are noun phrases, so that the system is designed to translate only the noun phrases. The verbs can be used in the forms of to + infinitive , verb-ing, and verb-ed. The system can not treat the embedded sentences which are introduced by relative pronouns.

Then the essential structures the system can treat are composed of simple noun phrases, verbs of the forms of to-infinitive, verb-ing, and verb-ed, and prepositions. Here a simple noun phrase means the juxtaposition (endocentric structure) of adjectives, nouns, and some other elements. The word order of a simple noun phrase can be the same in English and Japanese. The sentential structures obtained after parsing each simple noun phrase into a noun is called a skeleton pattern. We can expect that the variety of such skeleton patterns is very few for the restricted area of titles of scientific and engineering papers.

When the variety is very few, we do not need further syntactic analysis for these skeleton patterns. For each skeleton pattern the corresponding Japanese skeleton pattern (word order change) can be given. Thus the subgrammar in this system is a very peculiar one which is an accumulation of heuristics of the title structures. We utilized this specific nature of the titles in our machine translation system.

The correct translation rate for the wide variety of scientific and engineering papers is about 80%, but for the specific areas of physics and mathematics from INSPEC database the score was about 93%. The system is now used for the conversion

of English databases into Japanese databases. This system thus opened a way for the Japanese people to make access to English databases in their own language.

2. SPECIAL CHARACTERISTICS OF TITLE SENTENCES

Title sentences of scientific and engineering papers in English have the following properties from the point of view of translation.

(1) Nouns in the titles are usually specific terminological words in a particular field. The translation of these words into Japanese is almost unique. This makes avoid a difficult problem of the selection of proper translation words from several candidates, which we encounter in ordinary words.

(2) Many colloquial expressions exist in the titles. These are regarded as idioms, and their internal structures are not analyzed. The whole expressions are stored in a dictionary with their Japanese translations.

(3) A simple noun phrase in English can be translated into Japanese by replacing each word into Japanese without any word order change.

(4) Many of the special terminological words in science and engineering are compound words. They are treated as such in a dictionary. When the translation of a simple noun phrase according to (3) is not acceptable, the phrase is stored in a dictionary as a compound word with its translation. Therefore the dictionary look-up is done by the longest match principle.

(5) The word order change in the translation is only possible in the cases where verbs and prepositions are used. This word order change can be done at the level of skeleton patterns.

3. DICTIONARY LOOK-UP

The block diagram of our title translation system is shown in Fig. 1. The first step is the dictionary look-up of words and idioms. We gathered a lot of specific expressions as idioms, such as "time varying (mechanism)", "based on ...", and so on. "verb-ing" can be a noun, adjective, and present participle, but there are many verb-ing's whose grammatical function is almost unique: accounting, bonding, engineering and so on as nouns, superconducting as adjective, and using, determining as verbs which demand objects or complements. The dictionary has this information.

4. CONJUNCTIVE PHRASE

The second step is the parsing of conjunctive phrases by "and" and "or". As is well known there is an ambiguity for the conjunctive phrases of the forms:

A and B of C ,

Adjective + noun + and + noun,

and so on. It is very difficult to determine the scope of conjunctive phrases, and to get the correct parsing without the detailed semantic analysis. The present program parses simply the nearest two terms which have the same parts of speech, such as:

adj. + and + adj. ⟶ adj.

verb + and + verb ⟶ verb

verb-ing(-ed) + and + verb-ing(-ed) ⟶ verb-ing(-ed)

noun + and + noun ⟶ noun

Special consideration is given to the following specific conjunctive phrase:

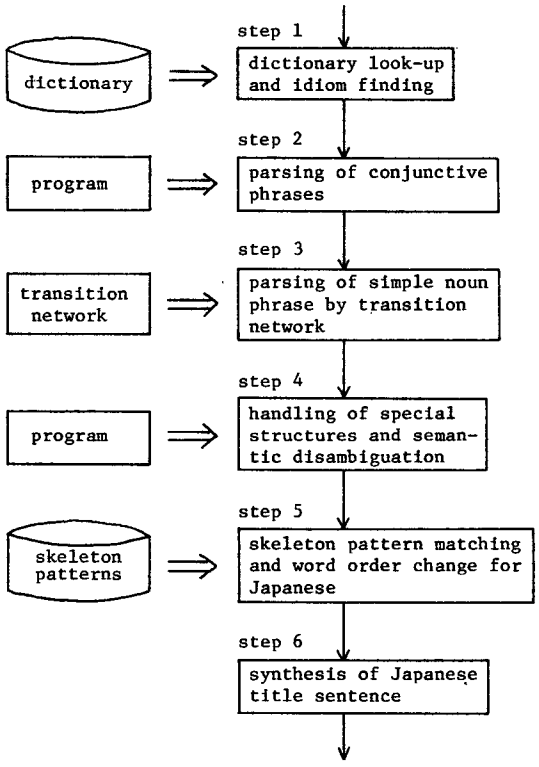

Fig. 1. Flow of Title Translation.

prep. + noun + and + prep. + noun ⟶ (prep. + noun) + and + (prep. + noun)
⟶ prep. + noun

Conjunctive structures such as,

(noun + prep. + noun) + and + (noun + prep. + noun)

(adj. + noun) + and + noun

can not be analyzed correctly.

5. SIMPLE NOUN PHRASE

Next step is the parsing of a simple noun phrase, which may include some other parts of speeches. The recognition of a simple noun phrase is done by the finite automaton model shown in Fig. 2. The recognition starts from the initial state, and the proper transfer of the state is done for the sequential input of words. When the automaton reaches to the final state the recognition of the end of a simple noun phrase is ended. The word order of the corresponding Japanese is the same as English within the scope of a simple noun phrase.

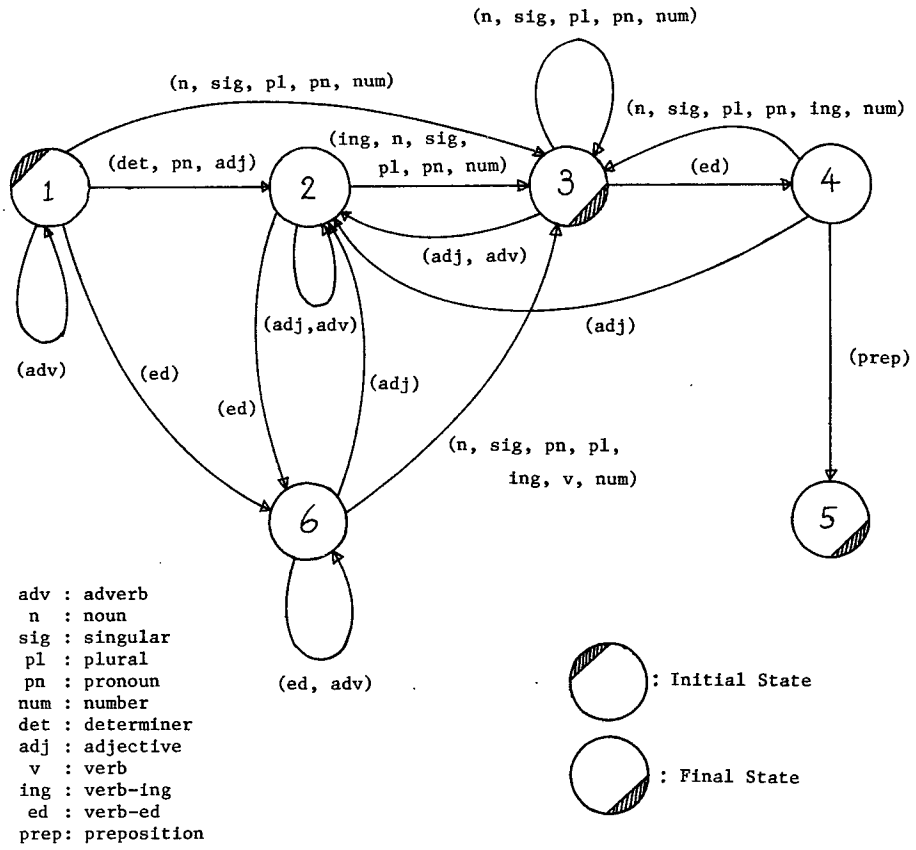

Fig. 2. Transition Network (TN).

6. SPECIAL WORD SEQUENCE

There are some particular word sequences which must be treated separately. Typical ones are as follows.

(a) n_1 + of + n_2 : This word sequence is regarded as a noun after parsing. This is translated into the Japanese word order : n_2 + の + n_1.

(b) prep. + n (at the beginning of the titles) : An example is "On pattern recognition". In this case, very tricky treatment is done as "prep. + n ⟶ n". This means that prep. is an accessory to the noun phrase (n) which follows it, and the structure of this noun phrase is the main part of the analysis. The translation is first done to the noun phrase, and at the final stage the translation of the preposition is attached to the end of the translated noun phrase.

(c) verb-ed + prep. : This structure is just parsed to prep. which has the modifying term of verb-ed. The Japanese translation is "prep. + verb-ed + された (passive particle). An example is :

a paper presented to a conference ⟶ a paper to a conference
(presented)

⟶ 会議 (conference) へ (to) 提出さ (present) れた (passive) 論文 (paper)

(d) verb-ing + prep. (at the beginning of the titles). An example is "concerning to ...". Syntactically "verb-ing" in this case plays the same role as a noun. So it is replaced by noun.

(e) noun + verb-ing + prep. : The determination of the grammatical role of verb-ing in this case is very difficult. By the title sentences it is frequent that verb-ing is used as a noun (gerund), and the interpretation of

noun + verb-ing ⟶ noun + noun ⟶ noun

is adopted.

7. SEMANTIC DISAMBIGUATION

After the parsing of the above particular structures there still remain some more difficult structures which require semantic treatment. "verb-ing + noun" is a typical such structure. Verb-ing can be either a modifying element to the noun, or a present participle which requires the noun as an object. An example is:

measuring temperature ⟶ 温度 (temperature) 計測 (measuring)

measuring device ⟶ 測定 (measuring) 器具 (device)

Therefore the check must be done between the verb and the noun which follows as to whether the noun can be a subject or an object to the verb.

For this purpose five semantic elements are introduced. These are shown in Table 1 with some nouns classified by these elements. The same semantic information is used to denote what kind of nouns can be a subject or an object to a verb. For example the subject nouns to the verb "measure" have the semantic categories of tool and theory, and the object nouns for the verb have the semantic categories of physical object, and aspect. By checking these semantic relations the syntactic structure and the translation word order are determined.

verb-ing + noun (subject) ⟶ verb する noun

verb-ing + noun (object) ⟶ noun を verb する ...

Table 1. Semantic elements

tool : instrument, machine, probe, etc.
aspect : velocity, temperature, resistance, etc.
physical object : metal, water, oil, waveguide, etc.
theory : principle, technique, approach, etc.
unit : cm, degree, etc.

Such semantic checking is performed in the following syntactic structures.

(1) n + verb-ing : if semantic check does not work, verb-ing is regarded as a gerund and is modified by the noun.

(2) verb-ing + n : if the noun phrase (n) has an article, it is an object of the verb. If semantic check does not work, n is regarded as an object.

(3) n_1 + verb-ing + n_2 : Semantic check between the verb and n_1, and the verb and n_2 is done. If semantic check does not work, the interpretation is that n_1 is an object of the verb, and verb-ing modifies n_2.

(4) prep. + verb-ing +prep. : verb-ing is understood as a gerund.

Table 2. Skeleton patterns and the frequency of their usage in INSPEC translation.

English skeleton pattern	Japanese word order	Frequency for INSPEC titles
(1) -ing · n	-ing · n	0
(2) n	n	466
(3) n · -ing	n · -ing	0
(4) n_1·prep·n_2	n_2·prep·n_1	536
(5) n_1·prep·n_2·ing	n_2·ing·prep·n_1	0
(6) n_1·$prep_1$·n_2·ing·$prep_2$·n_3	n_3·$prep_2$·n_2·ing·$prep_1$·n_1	0
(7) n_1·$prep_1$·n_2·$prep_2$·n_3	n_3·$prep_2$·n_2·$prep_1$·n_1	147
(8) n_1·$prep_1$·n_2·$prep_2$·n_3·$prep_3$·n_4	n_4·$prep_3$·n_3·$prep_2$·n_2·$prep_1$·n_1	32
(9) n_1·$prep_1$·n_2·$prep_2$·n_3·$prep_3$·n_4·$prep_4$·n_5	n_5·$prep_4$·n_4·$prep_3$·n_3·$prep_2$·n_2·$prep_1$·n_1	2
(10) n_1·$prep_1$·n_2·$prep_2$·n_3·$prep_3$·n_4·$prep_4$·n_5·$prep_5$·n_6	n_6·$prep_5$·n_5·$prep_4$·n_4·$prep_3$·n_3·$prep_2$·n_2·$prep_1$·n_1	0
(11) n_1·prep·n_2·v·n_3	n_2·prep·n_1·は·n_3·を·v	1
(12) n · v · adj	n·は·adj·v	0
(13) n_1 · v · n_2	n_1·は·n_2·を·v	1
(14) n_1·v·n_2·prep·n_3	n_1·は·n_3·prep·n_2·v	1
(15) n_1·v·prep·n_2	n_1·は·n_2·prep·v	0
(16) v · n	n · v	2
(17) v·n_1·n_2	n_1·は·n_2·v·か	1
(18) v·n_1·prep·n_2	n_2·prep·n_1·を·v	1

8. SKELETON PATTERN

The parsing process thus far produces a skeleton pattern for each title sentence. For example:

\# An Automated General Purpose Test System for Solid State Oscillators.

(Skeleton) System for Oscillators (n + prep. + n)

\# A Laser Doppler Technique for Measuring Flow Velocities in High Current Arc Discharge.

(Skeleton) Technique for Measuring Velocities in Discharge.
(n + prep. + ver-ing + n + prep. + n)

The skeleton patterns obtained from ten thousand title sentences are astonishingly few. These are shown in Table 2, with the frequency of occurrence of each pattern for about a thousand title sentences of physics and mathematics in INSPEC database. The Japanese word order is also given to each skeleton patterns.

The translation of prepositions is set unique by the present program as shown in Table 3. There are of course several cases where different Japanese expressions should be adopted for a preposition depending on the context. This is an important problem to be solved in the future.

Table 3. Translation of preposition.

of	の	to	への (n・to・n)
by	による	on	についての
with	による	in	での
at	における	about	について
for	のための		

9 TEST RESULT

A test result of the title translation from INSPEC database is shown in Table 4. Average time necessary for the translation of a title is 0.1 second. After the translation of 1000 titles, the dictionary was updated by the new words which appeared in the input data and which were absent in the dictionary. Then the same 1000 titles were again translated, and the rejection was checked. Then the next 1000 titles were handled in the same way, and so on.

Table 4. Test result of title translation from INSPEC database.

title nimber	computer time(sec.)	rejected title sentences	unregistered new words	sentences translated after the new word registration	untranslatable sentences after the word registration
1 ~ 1000	100.16	49	816	38	11
1001 ~ 2000	107.54	39	567	23	16
2001 ~ 3000	115.4	29	479	14	15

Computer used is M200 (one of the biggest computers in Japan).

With 3000 titles from INSPEC the rejected were only 42 titles (1.4%). Many of the rejected titles had the structures which the system can not accept, such as normal sentential structures, and question forms. The system can only accept the noun phrases without any embedded sentential structures by relative pronouns.

Among the translated titles, about 5% were wrong or ununderstandable. Many of these errors came from the wrong parsing of conjunctive phrases. Some examples of the translation are shown in Table 5. For some other databases in English the correct translation rate was about 80%. This rate depends heavily on the dictionary contents.

10. CONCLUSION

The translation system is now being used on trial basis at Tukuba Research Information Processing System (RIPS) of the Agency of Industrial Science and Technology. The titles, keywords, and some other journal information of INSPEC database are translated into Japanese, and a new database in Japanese language is created. Retrieval can be done by Japanese language by using Chinese characters and Kana letters to this database of INSPEC Japanese version.

The system seems to be practically usable, and the program is being transferred to a few other database centers for their use in the conversion of English database into Japanese database.

Table 5. Example of English Japanese translation.

THERMOHYDRAULIC ANALYSIS OF GAS-COOLED ROD ASSEMBLIES IN NUCLEAR REACTORS
核反応器での気体冷却捧組立の熱水圧解析

BEHAVIOR OF DRAG DISC TURBINE TRANSDUCERS IN STEADY-STATE TWO-PHASE FLOW
定常状態二位相フロウでのドラッグディスクタービン変換器の特性

VOID FRACTION CORRELATION OF TWO-PHASE FLOW OF LIQUID METALS IN TUBES
管での液体金属の二位相フロウのボイド分数相関

COMPARISON OF THE ORDER OF APPROXIMATION IN SEVERAL SPATIAL DIFFERENCE SCHEMES FOR THE DISCRETE-ORDINATES TRANSPORT EQUATION IN ONE-DIMENSIONAL PLANE GEOMETRY
一次元平面幾何での離散縦座標輸送方程式のための数種空間差分方式での近似の次数の比較

GENERALIZED QUASI-STATIC METHOD FOR NUCLEAR REACTOR SPACE-TIME KINETICS
核反応器時空間動力学のための一般化準静的手法

SEMICLASSICAL CONVERGENT CALCULATIONS FOR THE ELECTRON-IMPACT BROADENING AND SHIFT OF SOME LINES OF NEUTRAL HELIUM IN A HOT PLASMA
熱プラズマでの中性ヘリウムのある線の電子衝撃広がり及びシフトのための半古典的収束計算

TRANSITION PROBABILITIES AND THEIR ACCURACY
遷移確率及びそれらの正確さ

THEORY OF RESONANCE-RADIATION PRESSURE
共鳴放射圧力の理論

EXCHANGED MOMENTUM BETWEEN A SURFACE WAVE AND ATOMS
表面波及び原子の間の交換した運動量