Concept and Structure of Semantic Markers for Machine Translation
in Mu-Project

Yoshiyuki Sakamoto, Tetsuya Ishikawa, Masayuki Satoh

Electrotechnical Laboratory, Univ. of Library & Information Science, Center of Science and Technology
Nihari-gum, Yatabe schi, Nagaotsu Chiyoda-ku
Ibaraki, Japan

O. Abstract

This paper discusses the semantic features of nouns classified into categories in Japanese-to-English translation, and proposes a system for semantic markers. The system syntactic analysis is carried out by checking the semantic compatibility between words and nouns. The semantic structure of a sentence can be extracted at the same time as its syntactic analysis.

We also use semantic markers to select words in the transfer phase for translation into English.

The system of the Semantic Markers for Nouns consists of 13 conceptual facets including one facet for 'Others' (discussed later), and is made up of 49 fillial slots (semantic markers) as terminals. We have tested about 3,000 sample abstracts in science and technological fields. Our research has revealed that our method is extremely effective in determining the meanings of Japanese verbs (basic Japanese verbs) which have broader concepts like English verbs, 'make', 'get', 'take', 'put', etc.

1. Introduction

Semantic features are introduced to ensure the maximum possible accuracy of syntactic analysis, transfer and generation. We aim at a well-balanced usage of syntax and semantics throughout the whole process of machine translation.

The present paper introduces semantic concepts for nouns classified according to facets and slots which we called semantic markers. Then we show how these semantic markers are written in the respective lexicons for analysis, transfer and generation, and how effective they are in improving the quality and accuracy of the machine translation system in each phase of analysis and transfer. Therefore, semantic features are analyzed by the structure embedded into the case frame in Japanese syntax analysis: these features play an important role when selecting words in the transfer phase from Japanese into English.

Semantic features are more word specific. Pairs of deep cases and nouns should be written in the lexicon. However, due to the huge number of nouns, it is more effective to include pairs of the deep cases and semantic markers in the lexicon instead of nouns.

The Mu-project is a Japanese national project supported by the STA (Science and Technology Agency) "Research on a Machine Translation System (Japanese - English) for Scientific and Technological Documents."*

2. Transfer Approach to Machine Translation

We are currently restricting the domain of translation to abstract papers in scientific and technological fields. The system is based on a transfer approach and consists of three phases: analysis, transfer and generation.

In the first phase, morphological analysis divides sentence into lexical items, then syntactic analysis is carried out by syntactic and semantic analyses of Case Grammar in Japanese. In the second transfer phase, lexical features are transferred and at the same time, the syntactic structures are transferred by making them match tree patterns between Japanese and English. Here, we use semantic features to select words for translation into English. In the final generation phase, syntactic structures are generated by the Phrase Structure Grammar and the morphological features of English.

The following describes the processing functions employed in our system.

Morphological analysis and generation program are described in LISP, which is adequate for morphological process in Japanese and English, while syntactic analysis, transfer and generation programs are written in GRADE (Grammar Describer). Such process written in GRADE are independent of natural languages in machine translation. GRADE allows a grammar writer to write grammars using the same expression in all three phases.

Grammatical rules written in GRADE (Grammar Describer) are translated into internal forms, which are expressed by S-expression in LISP. This translation is performed by GRADE translator.

3. Concept of a Dependency Structure based on Case Grammar in Japanese

In Japan, we have come to the conclusion that case grammar is the most effective one for Japanese syntactic analysis in machine translation systems. This type of grammar has been proposed and studied by Japanese linguists before Fillmore's presentation.

As the word order is heavily restricted in English syntax, ATNG (Augmented Transition Network Grammar) based on CFG (Context Free Grammar) is adequate for syntactic analysis. However, Japanese word order is almost unrestricted and postpositional case particles play an important role as deep cases in Japanese sentences. Therefore, case grammar is the most effective method for Japanese

* This project was carried out with the aid of a special grant for the promotion of science and technology from the Science and Technology Agency of the Japanese Government.
syntactic and semantic analyses.

In Japanese syntactic structure, the word order is unrestricted except for predicates (verbs or verb phrases) which will be located at the end of sentences. In Case Grammar, verbs play a very important role in syntactic analysis, and the other parts of speech act only in partnership with or subordinate to verbs. That is, syntactic analysis is made by checking the semantic compatibility between verbs and nouns. Consequently, the semantic structure of a sentence can be extracted at the same time as syntactic analysis.

1. Morphological Analysis: Segmentation of a Japanese sentence by Lexicon Database

Ex. Input sentence: "我々は計算機能を文書に翻訳した。" is segmented as follows:

2. Syntactic Analysis: The item-to-item relationship of the sentence is analyzed to give syntactic features for the respective items.

3. Lexical features are transferred and the syntactic structure is transferred by matching patterns between Japanese and English.

4. Syntactic generation: The word order in English is converted according to Phrase Structure Grammar.

5. Morphological generation: Inflectional features such as tense, aspect, etc. are attached.

---

Table 3.1. Case Labels for Case Frames

| Japanese Label | English | Examples |
|---------------|---------|----------|
| (1) 主語 | Subject | ～が |
| (2) 対象 | Object | ～を |
| (3) 受け手 | Recipient | ～を与える |
| (4) 与え手 | Giver | ～から受ける、奪う |
| (5) 相手1 | PARTner | ～と協議する、異なる |
| (6) 相手2 | OPponent | ～から無理する、独立する |
| (7) 時 | Time | 1980年 |
| (8) 時-始点 | Time-FROM | 5月から |
| (9) 時-終点 | Time-TO | 来年まで |
| (10) 時間 | DURATION | 5分間加熱する |
| (11) 場所 | SPACE | ～に位置する、～で発生する |
| (12) 場所-始点 | Space-FROM | ～から帰る |
| (13) 場所-終点 | Space-TO | ～へ送る、～に到達する |
| (14) 場所-経過 | Space-THROUGH | ～を経て、上空を飛ぶ |
| (15) 始状態 | STATE | ～から5.5%へ引き上げる |
| (16) 終状態 | GOAL | ～から日本語に翻訳する |
| (17) 属性 | ATTRIBUTE | 流通性に富む、欠ける、之を |
| (18) 原因・理由 | CAUSE | 事故で死ぬ、～から分かる |
| (19) 手段・道具 | TOOL | オペレータで、ドリルで |
| (20) 材料 | MATERIAL | ターボエンジンで |
| (21) 構成要素 | COMPONENT | ～から成る、～で構成する |
| (22) 方式 | MANNER | 並列に、10m/secで |
| (23) 条件 | CONDITION | 焦点速度で決まる |
| (24) 目的 | PURPOSE | ～に適用する、〜を必要として |
| (25) 役割 | ROLE | 議員に選ぶ、〜として用いる |
| (26) 内容規定 | CONTENT | 〜と呼ばれる、〜が準備 |
| (27) 順序規定 | RANGE | 〜に次って、〜に関連して |
| (28) 順序 | TOPIC | 〜は、〜とは |
| (29) 視点 | VIEWpoint | 〜から立場 |
| (30) 比較 | COMPARISON | 〜より大きい、〜に劣る |
| (31) 従属 | DEPEND | 〜とともに、〜に伴随 |
| (32) 集合 | GROUP | 〜％増加する、3℃以下に |
| (33) 質問 | QUESTION | ～である |
| (34) その他 | ETC | ～ |

Note: The capitalized letters are used as abbreviations.

To write the semantic markers for nouns in the case frame of the verb lexicon, reference is made to the noun lexicon for these nouns.

Note that we write only the semantic markers for these nouns in the context of our samples. *Kouju-shi* as surface cases and case labels as deep cases are described for *Yogen*. Then semantic markers for nouns preceding to *Kouju-shi* are described.

5. Semantic Markers for Nouns

This section describes what the system of semantic information for nouns is and what the concept of semantic markers is and how semantic
markers are attached to nouns.

5.1 System of Semantical Information for Nouns

1) Study

In the primary stage of our study, we thought that all nouns were symbols to display the following concepts recognized by humans. We set up four concepts in highest level: 'Concrete objects', 'Abstract concepts', 'Phenomena', and 'Human actions'. Concrete objects are the selfsame objects in the world. Abstract concepts are the standards which fix intellectual activities of humankind. Phenomena include both social phenomena and natural phenomena. Human actions are the selfsame acted by humans. We assigned facets to these four concepts. Then we further extracted the feature of a part from these facets and assigned a new facet 'Parts'.

Similarly, another concept of 'Attribute' was extracted from 'Phenomena' and 'Human actions'. This feature is crucial especially for action nouns. Thus we added two facets: 'Parts' and 'Attribute'. Nouns also include concepts of measurement, space & topography and time. So we added three facets: 'Measurements', 'Spaces & Topography', and 'Time'.

We classified into more concepts as follows:

The concept of concrete objects are classified into 'Concrete objects', 'Animate objects' and 'Inanimate objects' which constitute three independent concepts. The concept of human actions was classified into two facets, 'Sense & Feeling' and 'Actions'.

We called the scope formed by the concept "conceptual category". It is difficult to define the conceptual scope explicitly. The concept which can be defined explicitly in the conceptual category is called a facet. The facet is subclassified into a number of semantic slots. This relation is illustrated in figure 5.1.

![Figure 5.1: Relationship between Facet and Semantic Markers](image)

2) Subclassification of Facets

Facets, for example, were subclassified into slots as follows, the facet of Animate objects was subclassified as semantic slots "humans", "animals", and "plants". The facet of Phenomena was subclassified as slots 'natural phenomena', 'physical phenomena', 'power and energies', 'physiological phenomena', 'social phenomena' and 'social systems and customs'. We then set up an 'others' slot in each facet, for these words which cannot be assigned to any slot. The use of these slots is explained in section 5.3. We will study 'others' slots through semantic analysis for nouns; new slots or facets may have to be assigned.

These semantic slots and facets are named semantic markers. The System of Semantic Markers for Nouns is shown in figure 5.2. The system of semantic markers for nouns is made up of 13 conceptual facets including 'Others' markers, and 49 filial slots as terminals.

We also use Special Semantic Markers for 'functional words' which represent some patterns, syntactic or semantic information. For example, the word 'comparison' presupposes more than two nouns (arguments): comparison between 'A' and 'B'. Then, 'WS(Relation)' as a special semantic marker is attached to the word 'comparison'. The word 'time' assumes time case. These features suggest an effective device for semantic analysis.

5.2 Concept of Semantic Markers

The following describes concepts of 12 facets in the System of Semantic Markers for Nouns ('Other' (ZZ) not included).

1) Nations and Organizations (OF)

This conceptual facet includes words related to such functional human groups as nations, parties, corporations and organizations. Words in this facet can occur with volitional verbs, when used as subjects.

2) Animate Objects (OV)

This conceptual facet includes such names as that of man, animal, and plant. However, names of organs of the animate objects are included in the slot of 'Organs or Components' (EL) under the facet of 'Parts'. Names of diseases are included in the slot of 'Physiological phenomena' (PO) under the facet of 'Phenomena'.

3) Inanimate Objects (IS)

This conceptual facet only includes words related to concrete objects in the animate objects, such as natural substances, parts and materials of products, artificial substances and institutions. The objects which do not exist as concrete objects are included in the facet of 'Intellectual Objects'.

4) Intellectual Objects (IO)

'IO' includes words related to theories, abstract tools and materials, intellectual products that are created by human intellectual activities.

5) Phenomena (PO)

'PO' includes words related to natural phenomena, physical phenomena, power & energies, physiological phenomena, social phenomena and systems/customs. Words having causal properties are attached to words under this facet using plus minus signs ('-' and '—'). Sign '+' denotes desirable conditions (e.g. success), while sign '—' indicates undesirable conditions (e.g. suicide).

6) Sense and Feeling (SO)

'SO' includes words related to human mental phenomena such as feeling, reaction, recognition and thinking.

7) Actions (AO)

'AO' includes words related to human activities such as human actions and movements.

8) Parts (CO)

'CO' includes such words related to parts and components of concrete objects as parts, components and organs.
Figure 5.2. System of Semantic Markers for Nouns
9) Attributes (AO)

"AO" includes words related to attributes of concrete objects and abstract concepts. Their slots consist of attribute's name, and attribute's values with causal relations, shapes, structures, constructions and nature.

For example, the word ‘color’ is the attribute's name (then, marker in AO), words such as “red” and “white” are attribute's value (AC).

10) Measurements (MO)

"MO" includes words related to numerals, name for numerals, standards, and units for measurement. Examples are "argument", "fee", "standards", and "kilometer".

11) Space and Topography (SA)

"SA" includes words referring to spatial extension of concrete objects and abstract concepts. Examples are direction, area, orbit and Brazil.

12) Time (TT)

"TT" includes words related to time points, time duration and time attributes, as ‘autumn’, ‘for a week’, ‘every day’ and ‘life time’.

5.3 How to attach semantic markers to words

The semantic markers for nouns are determined in the following steps.

1) Attach semantic markers to the following words:

Proper noun
Common noun
Action noun 1 (Shahen-mei shi)
Action noun 2 (except action noun 1)
Adverbial noun (only when the words include the concept of ‘Time’ or ‘Location’)
Interrogative pronoun
Personal pronoun
Demonstrative pronoun (only when the words include the concept of ‘Location’)

2) Attach semantic markers to the words according to the semantic scope and examples given in the 'definition table' of semantic markers.

3) Do not attach semantic markers to the following words:

Molecular formulas
Arithmetic expressions
Names of product models

4) If a word belongs to multiple slots in the same facet, attach all relevant markers.

5) If a word belongs to a facet but this word not belongs to any appropriate slot in the facet, attach ‘Others’ marker in this facet to that word.

6) If a word is equal to a facet name itself, attach the semantic marker of that facet name to the word.

7) If the concept of a word is not included in any facet, attach ‘Others’ facet ZZ to that word.

8) For compound words consisting of more than one word, attach the markers putting into consideration semantic information of the compound words themselves; do not always attach the marker only to the last element of the compound.

6. Semantic Information for Adverbs

In our system, adverbs are subclassified as follows.

1) Adverb of condition (Joukutsu fuku shi)

2) Adverb of degree (Teido fuku shi)

3) Adverb of statement (Chinjutsu fuku shi)

4) Adverb of quantity (Shuryou fuku shi)

Besides, the facets of verbs are classified: ‘Mood’, ‘Aspect’, ‘Tense’, and ‘Degree’. They contrast specific adverbs. Then semantic information for adverbs is used to ensure more accurate translations. Semantic information for adverbs are defined according to the concept aspects as follows.

1) Semantic information on mood determined by the adverb of statement

Subjunctive (e.g. if), Interrogation (e.g. when),Negation (e.g. not always), Possibility (e.g. possibly), Entireness (e.g. entirely), Concession (e.g. kindly).

2) Semantic information on aspects of verbs determined by the adverb of condition

Completion (e.g. finally), Progression (e.g. rapidly), Repetition (e.g. repeatedly), Concession (e.g. accordingly).

3) Semantic information on tense determined by adverb of condition and statement

Past (e.g. yesterday), Present (e.g. now), Future (e.g. tomorrow).

4) Semantic information on degree when the adverb or adjective can be modified by the adverb of degree and quantity

Scale (e.g. seriously), Degree (e.g. fairly).

7. Examples of Semantic Markers Used in Analysis

7.1 Determination of the Usage of Verbs by Case Patterns

Case patterns are used to determine the usage of verbs having broader concept. This is especially an effective method in determining the meanings of verbs (basic Japanese verbs) having broader concepts like English verbs, 'make', 'do', 'take', 'put', etc.

We take the word ‘hit’. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.

E.g. hit verb 'hit'. As an example and show the difference of the meanings of verb 'hit' by mean of case pattern (a), (b), (c). Furthermore we show the semantic markers which co-occur to each case.
In this way, we can determine the usage of verbs by means of case patterns and the semantic markers.

7.2 Interpretation of Optional Cases

One Kakuyo-shi (surface case) often plays the role of different deep cases in Japanese. Often, various optional cases are included in this deep case. Each optional case is determined by the combination of Kakuyo-shi (surface case) and the semantic marker of the noun which co-occurs with it in the dictionary. In the process of transfer, appropriate English prepositions must be specified according to the conditions of the optional deep case.

For example, take Kakuyo-shi "m". The optional case is determined from the semantic marker of noun and the semantic marker which co-occurs with the surface case of the verb. Then the case frame in English is selected and the preposition "in" is determined. This process is shown as follows.

Ex. (1) Source sentence: 主として現在市場で活躍している新興鉱物高級品があることが確認された。

Translated sentence: The newly discovered high-grade minerals are active mainly now in markets.

Ex. (2) Source sentence: 漬け物効果によるレーザ作動開始時における光電子エネルギーの低下を防ぐ方法を考案した。

Translated sentence: Problems are solved by two handling methods about laser absorption terms by the inverse damping radiation, and numerical solutions are obtained.

Explanation of Example (2): Using these Dictionaries, the noun "取り扱い法 (handling methods)" has semantic markers (IC and AN) while verb "解析 (solve)" has (DA, IT, IC or TS) for the noun preceding the Kakuyo-shi. "Handling methods" and "solve" match with each other with respect to "IC".

Example of selecting the translation for the Japanese verb, "含む" are as follows:

Ex. (1) Source sentence: タリウムおよび化カルシウムを含む高品質銅の放射および対対映示。

Translated sentence: The radiation and convection models of the vertical mercury arcs which contain the sodium and the scandium iodide.

Ex. (2) Source sentence: 与えられた正規方程式系を含む正規方程式系の構成法に関する具体的な構成法を示すことにより補練を図る。

Translated sentence: Lemmas are verified by showing specific constitution methods about constitution methods of the normal double orthogonal bases which include given normal double orthogonal systems.

Explanation of Example (1): Based on the Japanese to English Transfer Dictionary, Figure B.1, both "タリウム (sodium)" and "タリウムおよび化カルシウム (scandium iodide)" in Example (1) have OM. According to the conditions of the dictionary, OM matches one of the semantic markers (OM IP FN PB PP PE) in correspond to the appropriate case slot (in this case <市場 (object case)> ) of Japanese case frame in the dictionary for the verb "含む (contain, include)". So verb "contain" is selected.
3) Our system of semantic markers for nouns has been designed for Japanese nouns. We have to design a system of semantic markers for English nouns. Since recognition for its concept in English word is very difficult for the Japanese, we are also studying a machine translation project of the Science and Technology Agency.

4) When attaching semantic markers to verbs, we attached them without considering the relationship between noun and verb. That is, we attached them simply based on noun concepts. This is not adequate to handle nouns which are intrinsically related to verbs. One of the solutions to this problem will be to study the correlation matrix of the semantic markers for nouns in relation to the case frame of the verb.

5) Our concept system for words is not a static structure, but various semantic networks are constructed dynamically according to a given story. We must give thorough consideration to this prospective problems.

Acknowledgment

We would like to thank Prof. Makoto Nagao, Prof. Junichi Tsujii, Mr. Junichi Nakamura (Kyoto University), Mrs. Mutsumi Kimura (Institute for Behavior and Science), and Miss Masako Kume (Japan Convention Services, Inc.) and the other members of the Mu project working group for the useful discussions which led to many of the ideas presented in this paper.

References

In English

(1) Sakamoto, Y., Satoh, M. and Ishikawa, T.: Lexicon Features for Japanese Syntactic Analysis in Mu Project JF, COLING84, Stanford, 1984.

(2) Nagao, Makoto: Structural transformation in the generation stage of Mu Japanese to English machine translation system. Theoretical and methodological Issues in machine translation of natural language, Hamilton, New York, 1985.

(3) Tsujii, Jun-ichi: Science and Technology Agency's Machine Translation Project, Proceeding of the International Symposium on Machine Translation, JIPDEC, Tokyo, 1985.

In Japanese

(4) Nakai, H. and Satoh, M.; A Dictionary with Taigen as its Core, Working Group Report of Natural Language Processing in Information Processing Society of Japan, WGNL 30-7, July, 1993.

(5) Nagao, M.; Introduction to Mu Project, WGNL 30-2, 1993.

(6) Sakamoto, Y.: Yougen and Fuzoku-go Lexicon in Verbal Case Frame, WGNL 30-6, 1993.

(7) Ishikawa, T., Satoh, M. and Takai, S.; Semantical Function on Natural Language Processing, Proc. of 28th CIPSJ, 1984.

(8) Nagao, M., Tsujii, J., Nakamura, J., Sakamoto, Y., Torimura, T. and Satoh, M.: Outline of machine translation project of the Science and Technology Agency., Journal of Information Processing, Vol.28, No.10, 1985.