Object-Oriented Universal Grammar-Based Machine Translation (UGBMT)

Yukiko Sasaki Alam
Hosei University
Tokyo, Japan
sasaki@k.hosei.ac.jp

Abstract. This paper presents an object-oriented model for machine translation based on Universal Grammar, the Universal Lexicon and language-specific grammars and the lexicons, and demonstrates the internal structures of these linguistic constructs by following a step-by-step process of English to Japanese translation. It elucidates what elements are required in the Universal Lexicon and the lexicons of individual languages. The present model parses and generates sentences at three levels of structures: S-structure (Surface Structure), I-structure (Intermediary Structure) and U-Structure (Universal Structure). The present paper demonstrates the interaction of the three levels of structures in the process of translation, showing how economy and efficiency are achieved by incorporating the modules of Universal Grammar and the Universal Lexicon into the model of machine translation. This design makes each language grammar slim, distinguishing idiosyncrasies from elements of universal nature.

1 Introduction

This paper presents an object-oriented grammatical model for machine translation built on the assumption that there exists a component called Universal Grammar that contains linguistic information common to all languages and that language-specific grammars are composed of extensions of universal constructs as well as language-specific idiosyncrasies. Universal Grammar and the Universal Lexicon store language-independent information including:

- Universal Meanings that mediate translation at the deepest level;
- Semantic verb classes (which are sources of information on the aspects of events denoted by verbs and the semantic categories of the arguments of verbs);
- Semantic categories of words such as ANIMATE and HUMAN.

I am grateful for the helpful comments and suggestions provided by two anonymous refrees.

- Aspectually related Universal Meanings for verbs: for instance, the corresponding process verb meaning of SURPRISE is BECOME SURPRISED, and the corresponding state verb meaning is BE SURPRISED;

- Prototypical syntactic categories such as sentence, noun phrase, verb, and one-place verb.

Morphological information, on the other hand, is language-specific and stored in individual grammars. Word order is also language-specific and the information should be included in language-specific phrase structure rules. This paper demonstrates how language-independent and language-specific items of information are interwoven in the process of translation. The current model has the following characteristics:

- It is based on an object-oriented design;

- It consists of Universal Grammar, the Universal Lexicon and language-specific grammars and the lexicons;

- It divides sentence representation into three levels of structures (i.e.three levels of sentence understanding): language-specific S-structure (Surface Structure), I-structure (Intermediary Structure) and language-independent U-structure (Universal Structure);

- S-structure is composed of syntactic and functional categories and surface forms of words and morphemes while I-structure is made up of functional categories and Universal Meanings. U-structure includes information on semantic categories and semantic relations, in addition to those linguistic constructs at I-structure.

The proposed model is one of a very few machine translation models designed on an object-oriented architecture, although there have been not a few publications on object-oriented natural language processing (Li and Byant 1998, Lavoie, Rambow and Reiter 1997, Neuhaus and Hahn 1996, Saint-Dizier 1994, Seligman 1991, Miyoshi and Furukawa 1985, to name a few).

2 The Modules

This model is designed to translate from English to Japanese and vice versa. It is composed of three main Java packages, *universalgrammar*, *englishgrammar* and *japanesegrammar*.

The package *universalgrammar* contains a subordinate package, *verbclasses*, consisting of Java classes representing semantic verb classes such as *Amuse* verbs and verbs of change of possession. [1] It also includes a package entitled *semanticcategories*, which holds classes representing semantic categories. It houses a class entitled *UniversalLexicon* that holds lexical entries for Universal Meanings. A lexical entry for a Universal Meaning contains information, for instance, on the verb class when it is for a verbal meaning or information on the semantic category when it is a nominal meaning. It also contains a package entitled *syntacticcategories* representing prototypical syntactic categories such as sentence, noun phrase, noun, verb and one-place verb.

The packages *englishgrammar* and *japanesegrammar* each contain Java classes representing its own syntactic categories such as *EnglishSentence* and *JapaneseSentence*, and the lexicons such as *EnglishLexicon* and *JapaneseLexicon*. They also include packages containing classes for their syntactic and morphological rules. Java classes representing language-specific syntactic categories, *EnglishSentence* and *EnglishNounPhrase*, for instance, inherit attributes and methods from their super classes, *Sentence* and *NounPhrase* in the *suniversalgrammar*. Thus, these classes list only idiosyncratic properties, highlighting the difference from the properties of universal nature.

To recapitulate, the lexical entry for a content word in an individual language contains a definition on (a) the Universal Meaning, (b) the syntactic category, (c) the morphological information, (d) the idiomatic uses and (e) the peculiarities of the word. It may list more than one definition with a different Universal Meaning, even though the two definitions may have the same syntactic category. In this sense, a language-specific lexicon is similar to a monolingual dictionary we use in daily life except for the following two main points. One difference is that meanings listed in lexical entries in the current model are Universal Meanings rather than explanations, while the other difference is that lexical entries in the Universal Lexicon include information on verb classes and semantic categories in order to provide lexical words in individual languages with semantic information. Information on the semantic categories of a noun such as ANIMATE and METAL is obtained via the Universal Meaning listed in the Universal Lexicon, and therefore it is not included in lexical entries of individual languages. Language-specific morphological information such as whether a noun is countable and whether it is singular must be included in lexical entries for nouns in an individual language. As a result, this design reduces redundancy and brings to light what is language-independent and what is not.

The following chart shows the general organization of the current model:

[1] The terms *Amuse* Verbs and verbs of change of possession are from Levin (1993).

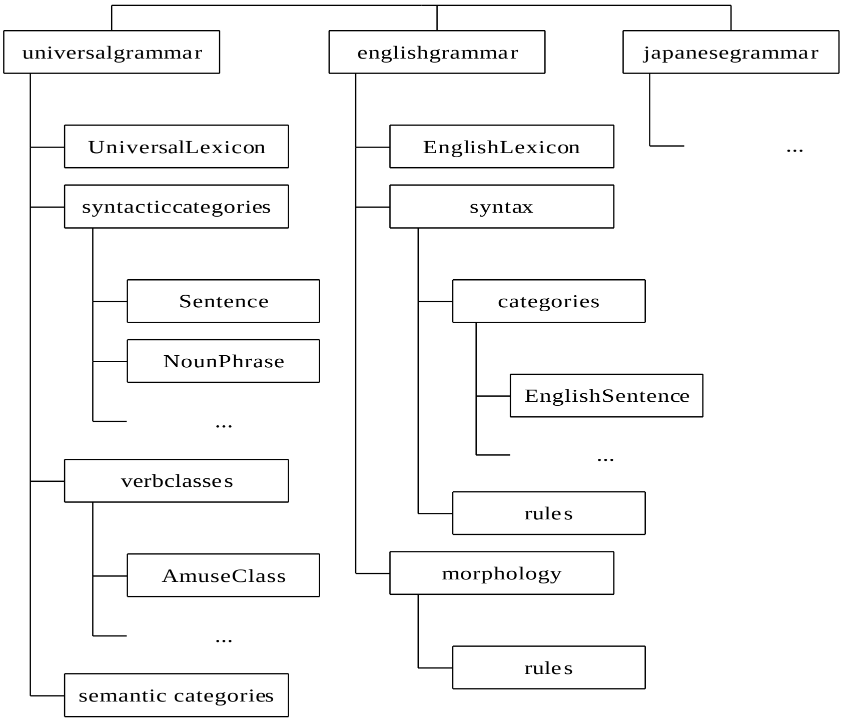

Fig. 1. The Modules[2]

3 Process of Translation

This model is designed to translate English into Japanese and vice versa. It is composed of three main packages, *universalgrammar*, *englishgrammar* and *japanesegrammar*, respectively representing Universal Grammar and English and Japanese grammars.

3.1 S-structure (Surface Structure)

An S-structure is composed of syntactic categories such as noun phrase, verb phrase and noun as well as functional categories such as head, complement and modifier. A prototypical sentence consists of a specifier noun phrase, a head inflectional element

[2] The names in lowercase represent Java packages while those in a mixture of uppercase and lowercase indicate Java classes.

and a complement verb phrase[3]. The Java class *Sentence* in *syntacticcategories* of *universalgrammar* is illustrated below:

Sentence
lexicon, sentence, modifier, specifier, subject, head, complement
(setters and getters) …

Fig. 2. The Java class *Sentence* in the package *universalgrammar*

As the specifier of a sentence is the subject, it is so defined in the class.

The class *EnglishSentence* extends *Sentence*, thus inheriting attributes and methods from the super class in *universalgrammar*. This relation of inheritance is able to highlight shared properties while helping avoid redundancy at the same time.

EnglishSentence
englishlexicon, englishsentence
specHeadAgreement …

Fig. 3. The Java class *EnglishSentence* in the package *englishgrammar*

It should be noted that the method "specHeadAgreement" ensures the English agreement in number and person between the specifier subject noun phrase and the head inflectional element.

As shown below, the S-structure of *The decision did not surprise me* consists of a specifier, a head and a complement, while the complement breaks down further into several layers of components:

[3] Sentences and nouns are built on the following three-level XP rules: ${}_{XP}$[Specifier ${}_{X'}$ [X Complement] where X is a head element and X' is an intermediary phrase. On the other hand, verbal phrases such as verb phrases, aspectual phrases and negative phrases are constructed on the following two-level XP rules: ${}_{XP}$[X Complement] where X is a head element.

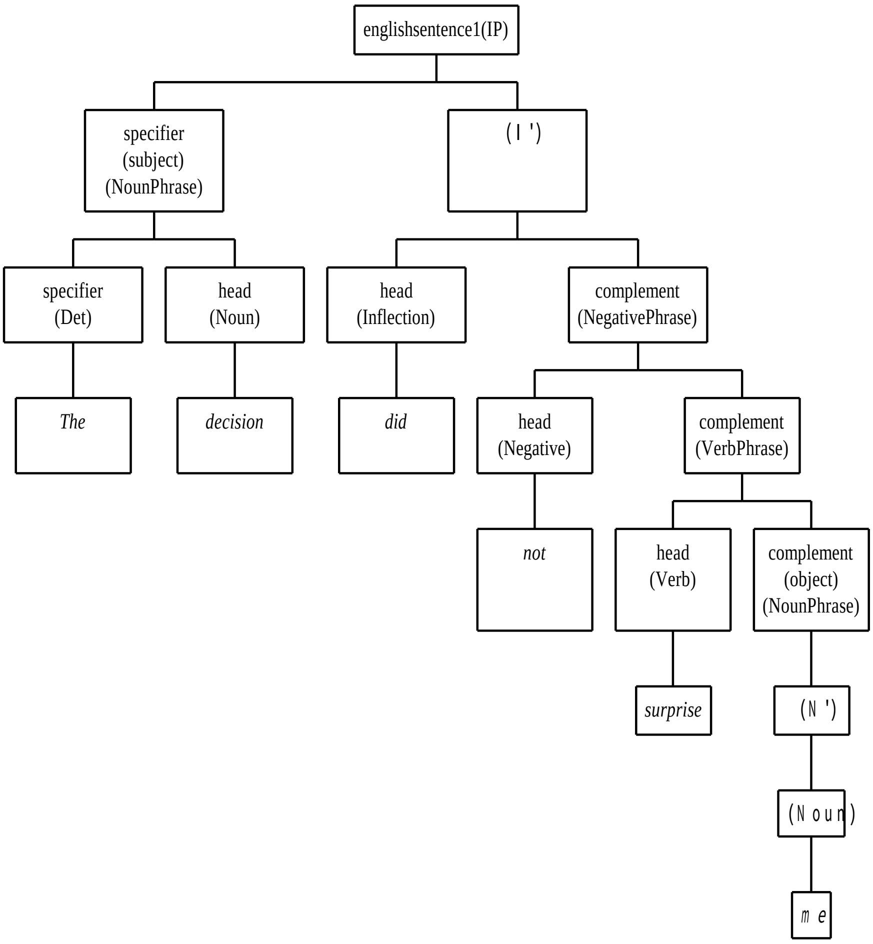

Fig. 4. S-structure for *The decision did not surprise me*.[4]

The above S-structure is composed of such functional categories as specifiers, heads and complements with labels of syntactic categories in parentheses. A simple sentence is an IP (Inflectional Phrase). The idea is that a proposition becomes a sentence once it is anchored with an inflectional element in this spatio-temporal world. The value of *englishsentence1.specifier.head* is the string *decision*, and the value of *englishsentence1.head* is the string *did*. The value of *englishsentence1.complement* is a NegativePhrase.

[4] The actual names of classes in Java for the English syntactic categories contain *English* at the beginning, like *EnglishNounPhrase* and *EnglishInflectionalPhrase*, but for the sake of space, the name of the language is omitted from the tree diagrams in this paper. This policy applies to classes for Japanese syntactic categories as well

3.2 I-structure (Intermediary Structure)

The I-structure for a sentence contains information on Universal Meanings, in addition to functional and grammatical information of the language. It does not retain information on syntactic categories and surface forms of words and morphemes any longer. Universal Meanings are obtained via the lexical entries for surface forms of words and morphemes in individual languages, subsequently replacing the surface forms at S-structure. An example I-structure is illustrated below:

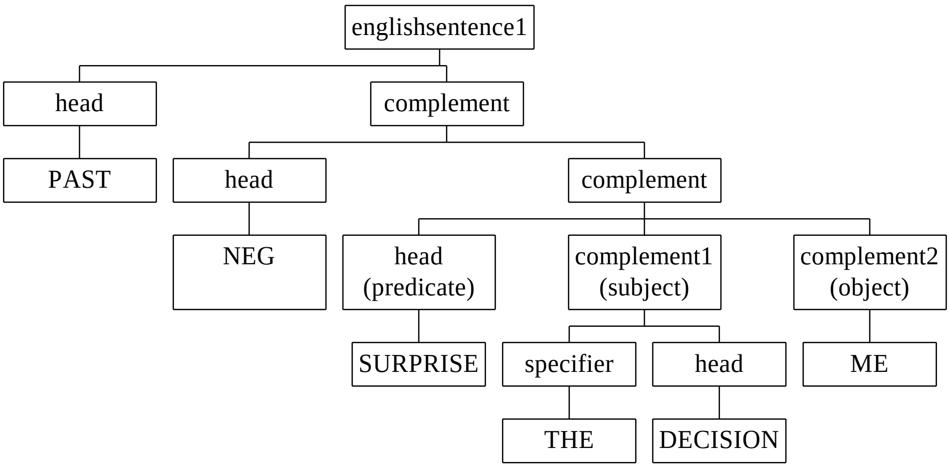

Fig. 5. I-structure for *The decision did not surprise me*

Note that the subject noun phrase of the sentence at S-structure is moved to a place closer to the predicate of the sentence, forming a natural group of complements of the predicate (SURPRISE). That is, at I-structure, the complements of a verb are reinstated at a verb phrase, which together compose a semantically complete unit. The verb meaning SURPRISE takes two complements or arguments, one that surprises someone and the other that gets surprised.

It should also be noted that the intermediary phrase IP’ at S-structure is absorbed into the sentence (IP) because it, without any functional role, is no longer required. At this level, any node only with syntactic information should be clipped.

3.3 U-structure (Universal Structure)

The U-structure holds information on event roles of the arguments of verbs such as AGENT, CAUSE and PATIENT, in addition to functional information and Universal Meanings inherited from the I-structure. For instance, the event role of the subject of *The decision did not surprise me* is CAUSE, and that of the object PATIENT. Where is such information derived from? It is obtained via the lexical entry for the Universal

Meaning SURPRISE in the *UniversalLexicon*. The entry for SURPRISE lists its verb class, which in this case is the class *AmuseClass*. This verb class contains information on the event roles of the arguments as well as the event aspect. All the verb meanings belonging to this class receive the same information, so that the lexical entry for each Universal Meaning belonging to the same verb class does not have to keep the same information individually. Following is the U-structure for *The decision did not surprised me*:

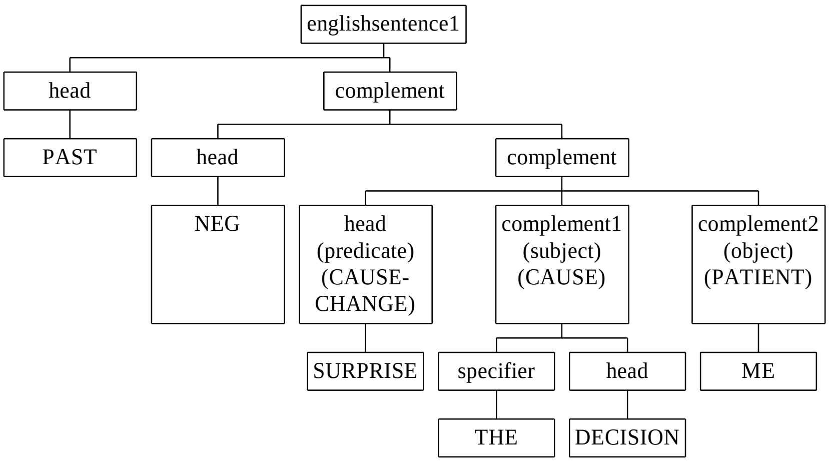
englishsentence1
head
complement
PAST
head
complement
NEG
head (predicate) (CAUSE-CHANGE)
complement1 (subject) (CAUSE)
complement2 (object) (PATIENT)
SURPRISE
specifier
head
ME
THE
DECISION

Fig. 6. U-structure for *The decision did not surprise me*

Although the event role of the subject of SURPRISE (a member of the *AmuseClass*) can be either CAUSE or AGENT, the event role CAUSE is assigned to the subject of the sentence, because the subject of SURPRISE, DECISION, is an inanimate entity, which rules out the possibility of being AGENT. The lexical entry for DECISION in the *UniversalLexicon* contains information on its semantic category.[5] The event role of the object of the *AmuseClass* is PATIENT, and the verb aspect is CAUSE-CHANGE. The values of the event roles and the verb aspect of SURPRISE, which are obtained from the *AmuseClass*, are placed in parentheses on Fig. 6 above.

3.4 U-structure to I-structure of the target language: grammatical demotion and promotion

In the process of transition from the U-structure to the I-structure for the target language, the grammar of the target language is consulted for well-formedness. In the case in question, the event role hierarchy in Japanese grammar is violated, because the

[5] Although no reference is made to ambiguity resolution in the current paper, it is facilitated by information on (a) the semantic categories of the arguments of verbs, (b) the aspect classes of verbs, and (c) the semantic categories of words, all obtained via *UniversalLexicon*.

event role CAUSE cannot obtain a higher grammatical status than PATIENT[6]. To satisfy this well-formedness condition, the grammatical function of the CAUSE argument of SURPRISE must be demoted to an oblique case, and consequently the grammatical function of the PATIENT argument is promoted to the subject. This operation results in a change in the aspect of the event: from the event that something surprises someone (CAUSE-CHANGE) to the event that someone gets surprised because of something (STATE-CHANGE). What is required after this change is a predicate with the PATIENT subject that still retains the meaning relating to SURPRISE. The right predicate is BECOME SURPRISED, the aspect of which is STATE-CHNAGE. In the *UniversalLexicon*, the determination of such predicates is automatic. The corresponding STATE-CHANGE predicate of SURPRISE is BECOME SURPRISED, and that of KILL is BECOME KILLED. Following is the Japanese sentence at I-structure after the change of grammatical statuses of the CAUSE and PATIENT arguments :

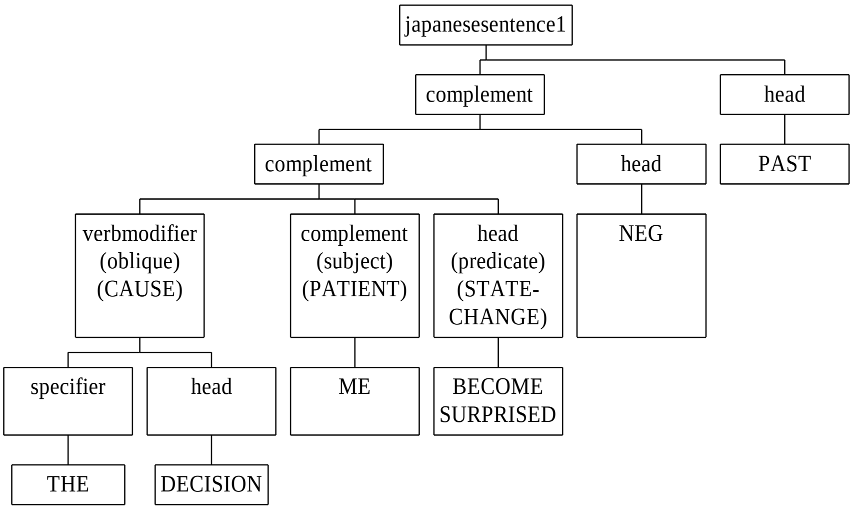

Fig. 7. I-structure for *watashi-wa sono kettei-ni odorokanakatta* ‘I did not get surprised at the decision.’

In Japanese, head elements consistently follow non-head elements such as complements and specifiers because, unlike English (a head-initial language), Japanese is a strict head-final language. Note that the grammatical category of the CAUSE argument is now an oblique, lower than the PATIENT subject, thus satisfying the well-formedness condition.

[6] The Japanese event role hierarchy and grammatical hierarchy are respectively: AGENT > PATIENT> ... > CAUSE and SUBJECT> OBJECT> ... > OBLIQUE.

3.5 S-structure for the corresponding sentence in the target language

At S-structure, Universal Meanings are replaced with surface forms of words and morphemes of the target language. Lexicons of individual languages also hold a list of pairs of both Universal Meanings and the corresponding surface forms. The replacement is carried out by consulting the list. The counterpart Japanese sentence at S-structure is illustrated below:

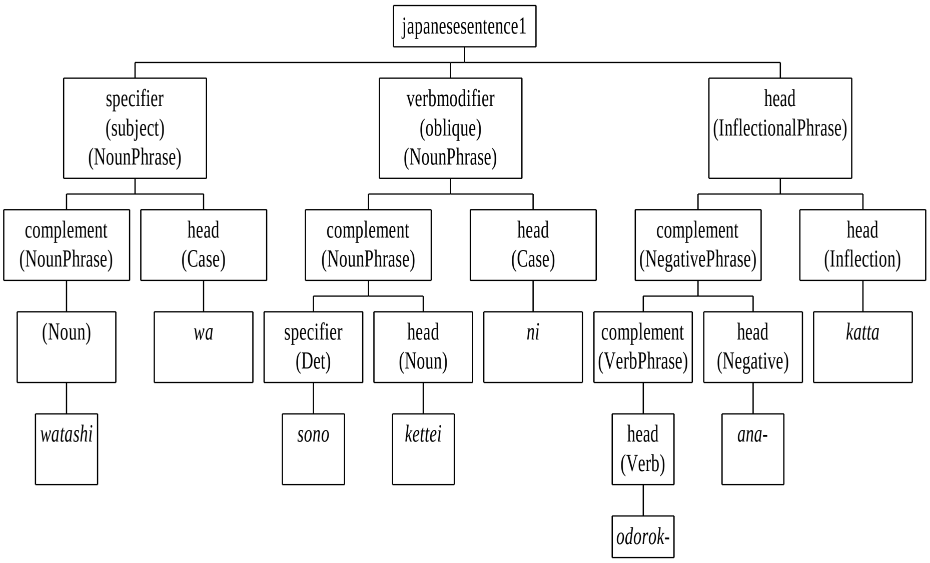

Fig. 8. S-structure for *watashi-wa sono kettei-ni odorokanakatta* ‘I did not get surprised at the decision.’

In the process of generating S-structures from I-structures, Japanese syntactic rules assign Case such as *wa* to noun phrases according to the discourse roles or grammatical roles they play. The complement and oblique arguments of the verb BECOME SURPRISED are moved up to the sentence (i.e. IP) level, leaving their head verb alone at the verb phrase component. The movement of verb complements to the sentence level at S-structure is attested by the free appearance of other elements such as time nouns and sentence adverbs before and after them. The order of the two elements in question is arbitrary as they appear in an arbitrary order in a real sentence.

3.6 Flow chart of the process of translation

The following chart shows the flow of the interaction at each level of structure with the grammars and lexicons.

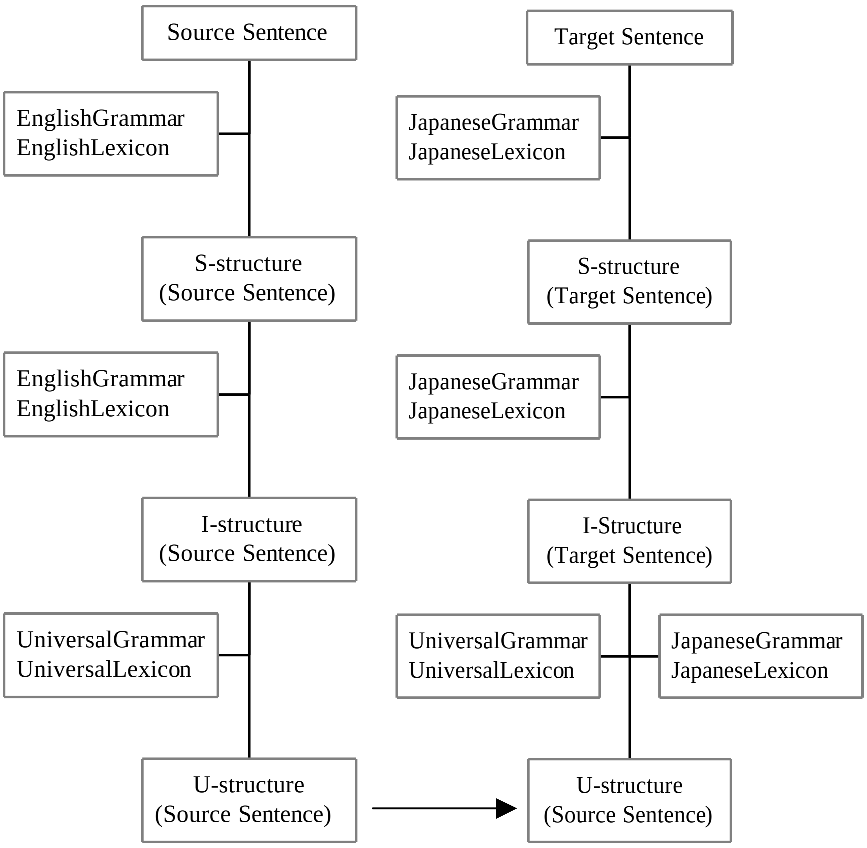

Fig. 9. Flow Chart of the Process of Translation

4 Conclusion

This paper has presented an object-oriented grammatical model for machine translation built on the assumption that there exists a component called the Universal Grammar containing linguistic information common to all languages. By incorporating prototypical syntactic categories, verb classes and semantic categories into Universal Grammar, and Universal Meanings into the Universal Lexicon, the current approach makes each language grammar slim, highlighting idiosyncrasies and universal properties. In addition, the three levels of sentence representation, S(urface)-structure, I(ntermediary)-structure and U(niversal)-structure, reflect differences in the levels of understanding of sentences. Finally, the proposed model offers a transparent organization of the modules, and offers an intuitively reasonable process of translation, which in turn facilitates the extensibility and modification of the model.

References

1. Levin, Beth: 1993, *English Verb Classes and Alternations*, Chicago: The University of Chicago Press
2. Lavoie, B., O. Rambow and E. Reiter. 1997, ‘Customizable descriptions of object-oriented models’, in *Proceedings of the Fifth Conference on Applied Natural Processing*, pp. 253-256.
3. Li, Li. & Barrett R. Bryant: 1998, ‘An efficient parsing model for unification categorial grammar with object-oriented knowledge representation and selection sets’, *International Jr. on Artificial Intelligence Tools* **7**,143-162
4. Miyoshi, Hideo and Koichi Furukawa: 1985, ‘Ronri programming gengo ESP ni-okeru Object shikoo koobun kaiseki (An Object-Oriented Parser in ESP, a Logic Programming Language)’, in Veronica Dahl and Patrick Saint-Dizier (eds) *Shizengo rikai to ronri programming*, Tokyo: Kindaikagakusha, pp. 68-89
5. Neuhaus, Peter and Udo Hahn: 1996, ‘Restricted parallelism in object-oriented lexical parsing’, in *COLING-96 Proceedings Vol. 1*, pp. 502-507
6. Saint-Dizier, Patrick: 1994, ‘Object-oriented logic programming for natural language processing’, in Patrick Saint-Dizier *Advanced logic programming for language processing*, London: Academic, pp. 211-252
7. Seligman, Mark. 1991, *Discourses from Networks Using an Inheritance-based Grammar*. Ph.D. thesis, Department of Linguistics, University of California at Berkeley.