HIERARCHICAL LEXICAL STRUCTURE AND INTERPRETIVE MAPPING IN MACHINE TRANSLATION

Teruko Mitamura
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213 USA
teruko@cs.cmu.edu

Eric H. Nyberg III
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213 USA
ehn@cs.cmu.edu

Abstract

Large-scale knowledge-based machine translation requires significant amounts of lexical knowledge in order to map syntactic structures to conceptual structures. This paper presents a framework in which lexical knowledge is separated into different levels of representation, which are arranged in a hierarchical model based on principles of knowledge representation and lexical semantics. The proposed methodology is language-independent, and has been used to organize lexical knowledge for both English and Japanese.

1 Introduction

The basic premise of knowledge-based machine translation is that accurate, high-quality translation requires a complete semantic interpretation of the input text (Carbonell and Tomita, 1987). Therefore, the analysis and generation components of a knowledge-based MT system must have at least the following functional parts: a grammar for the language, a lexicon for the language, a shared set of domain concepts, and rules that map syntactic structures onto semantic structures (or vice-versa for generation).

.The goal of our work has been to develop a methodology for the hierarchical organization of lexical knowledge (lexical entries and mapping rules) for knowledge-based MT (Goodman and Nirenburg, 1991; Mitamura, 1989). *Interpretive Mapping* refers to the relationship between predicate conceptual structures and syntactic structures, and involves two kinds of processes: one is a mapping between grammatical functions (e.g., subject, object) and semantic roles (e.g., agent, theme); the other is a mapping between words (e.g., *naguru* 'hit') and domain concepts (e.g., *HIT).

We have developed a shared hierarchical structure for lexical knowledge which can capture significant linguistic generalizations, eliminate redundancy, and facilitate both knowledge acquisition and efficient processing. We have implemented our hierarchy using FrameKit, an AI knowledge representation language that supports frames and multiple inheritance (Nyberg, 1988).

Our system demonstrates the integration of a linguistic formalism with a frame-based knowledge representation system. We have analyzed a large corpus of Japanese verbs and created a set of lexical frames, mapping rules, and an inheritance hierarchy for use in a working translation system.

2 Linguistic Motivation

Our methodology is based in part on recent work in lexical semantics (Jackendoff 1983, 1987; Levin, B. 1985, 1987, 1989; Hale and Keyser 1986; Fukui, Miyagawa, and Tenny 1985; Rappaport and Levin, B. 1986). The field of lexical semantics is concerned with the representation of syntactically relevant aspects of word meaning, especially the properties of argument-taking words like verbs.

Many researchers have noticed that semantically similar predicates tend to be syntactically similar, too. B. Levin (1987, 1989) examines many systematic semantic-syntactic correspondences, including linking regularities and transitivity alternations. *Linking* refers to associations between semantic arguments and grammatical relations. Common correspondences between semantic arguments and grammatical relations are called *linking regularities*.

2.1 Linking and Alternation

For example, in the causative use of *break* (e.g., *John broke the vase*), the subject *John* is linked to the agent semantic role, and the object *vase* maps to the theme semantic role. *Break* can be classified as a change-of-state verb, and the same pattern is observed in the causative use of other change-of-state verbs, such as *crack* and *melt*. Moreover, it is important to note that this pattern also holds for other classes of verbs (e.g., change-of-possession verbs like *give*).

It is also the case that the same verb can have more than one way of linking syntactic functions with semantic roles. These different linkings are called *valency alternations*, which include both transitivity alternations and alternate linkings of semantic arguments with syntax.

For example, *break* can also appear in sentences like *The vase broke*, where the verb assigns the theme semantic role to the syntactic subject. This is in contrast to the causative use of *break*, described above, where the verb assigns the agent semantic role to the syntactic subject and a theme semantic role to the syntactic object. This alternation, known as the Causative/Inchoative alternation, is also associated with change-of-position verbs like *drop* (*John dropped the ball* vs. *The ball dropped*), change-of-psychological-state verbs like *worry* (*John worried* vs. *Bill worried John*), etc. (B. Levin, 1989).

With some verbs, the mapping of one syntactic function may remain constant while others alternate. For example, in the sentence *John cut the meat*, the patient semantic role is assigned to the syntactic object; in *John cut at the meat*, the goal semantic role is assigned to the prepositional object. In both sentences, the agent semantic role is assigned to the syntactic subject.

Classes of verbs which undergo the same alternation tend to be semantically similar. Verbs like *hack* and *slash*, which belong to the same verb class as *cut*, undergo the same alternation mentioned above. However, semantically different verbs like *break* do not exhibit the same alternation:

```
(1) a.  He broke the cup.
    b. *He broke at the cup.
```

Linking regularities and transitivity alternations are used to identify the semantic roles of arguments and the semantic classes of verbs. That is, an argument which displays the same linking regularities as another argument might be assigned the same thematic role, and verbs which have the same transitivity alternations can be placed in the same class.

Transitivity alternations in English are marked in various ways. Many of them involve the alternation of an argument between object and prepositional phrase. In Japanese, however, valency alternations, (including transitivity alternations) are usually indicated by different case markers borne by the arguments of the verb. Every noun phrase in Japanese is marked postpositionally by a particle, such as *ga*, *o*, *ni*, and *de*. These markers indicate the case or other grammatical function of the nominals they are associated with.

For example, the *o/de* alternation appears with verbs like *oyogu* (swim), *sanposuru* (take a walk), and *hashiru* (run)[1].

```
(2) a. Taro ga kawa o oyoida
       'Taro swam down the river'
    b. Taro ga kawa de oyoida
       'Taro swam in the river'
```

2.2 Lexical Mapping

Another part of building a lexical semantic representation is to formulate links from lexical items to conceptual meanings; these links are called *lexical mappings*. Since the semantic properties of relations and objects (which are crucial in stating subcategorization restrictions) reside most naturally in a semantic domain model, it is necessary for a system to integrate the lexical level and the domain model so that semantic restrictions can be satisfied during parsing and generation.

In some cases, a lexical item may be linked to more than just a semantic head. For example, in the sentence *The pencil rolled off the table*, the meaning of *roll* must be represented by both a semantic head (e.g., *MOVE) and a semantic modifier indicating the manner of motion (e.g., (manner *ROTATION))[2]. As a result, lexical mapping may also require semantic feature assignment.

2.3 Summary

The motivation for our work has been the following set of observations, drawn from the linguistic phenomena mentioned in this section. An appropriate lexical representation must be able to represent the following:

- The linking of a particular syntactic function with a particular semantic role;
- A set of linking rules that indicate a particular alternation;
- A group of alternations that capture the general behavior of a class of verbs;
- An explicit representation of verb classes, to which particular lexical items may be linked;
- A set of lexical items, which contain both links to verb classes and links to semantic concepts in the domain conceptual hierarchy.

3 The Lexical Hierarchy

Our lexical hierarchy has five levels of representation, each corresponding to a linguistically meaningful unit of structure: (1) Mapping Rule Frames, which capture a particular correspondence between a syntactic function and a semantic role; (2) Mapping Pattern Frames, which capture a particular set of mapping rules, which correspond to one way of linking the arguments of a particular verb; (3) Mapping Type Frames, which capture the set of alternations (mapping patterns) allowed by a particular class of verbs; (4) Verb Class Frames, in which the generalization in verb linking behavior is captured; (5) Lexical Frames, in which particular lexical items (verbs) are represented as frames which are linked both to appropriate verb class frames and to conceptual frames in the domain concept hierarchy.

Figure 1 illustrates the inheritance relations between mapping rules, mapping patterns, mapping types, verb classes, and lexical frames in English.

3.1 Mapping Rule Frames

The mapping rule frames each map one grammatical function, such as subject or object, onto a semantic role, such as agent or theme. Each mapping rule is specified in a separate frame, as in the following:

```
a. (*agent-subj-mapping
     (:agent subj))
b. (*theme-obj-mapping
     (:theme obj))
c. (*theme-subj-mapping
     (:theme subj))
```

[1] For further detail and examples, see (Mitamura, 1989).

[2] This is similar to the notion of *conflation* discussed by Talmy (1985).

3.2 Mapping Pattern Frames

The mapping pattern frames represent particular bundles of mapping rules. For example, a mapping pattern frame which contains the agent-subject mapping and the theme-object mapping represents one mapping pattern, whereas a frame which contains just the theme-subject mapping represents another mapping pattern[3] (cf. Figure 1).

Syntactic constraint rules can be written in a mapping pattern frame to indicate that the associated mapping rules can apply only when these constraints are satisfied. Some examples of mapping pattern frames are shown below:

```
(*mapping-pattern1
  (syntactic-constraint
     (passive = -))
  (contain *theme-obj-mapping
           *agent-subj-mapping))
(*mapping-pattern2
  (syntactic-constraint
     (passive = -))
  (contain *theme-subj-mapping))
```

The frame *mapping-pattern1 captures one way of mapping the syntactic argument of a verb. The subject is mapped to the semantic agent and the object is mapped to the semantic theme. The *mapping-pattern2 frame indicates a mapping where the verb has one argument, the subject, and maps the subject to the semantic agent.

3.3 Mapping Type Frames

Mapping type frames contain sets of mapping rule patterns, and have the ability to capture both transitivity alternations in English and case alternations in Japanese (Mitamura, 1989). The two mapping patterns we mentioned earlier, 1) the agent-subject and the theme-object mapping, and 2) the theme-subject mapping, can be generalized as the causative-inchoative verb mapping type. In Figure 1, the causative-inchoative alternation is represented by *causative-inchoative. In Japanese, the alternation between an oblique argument with particle *o* and an oblique argument with particle *de* is captured by *obl-o/obl-de.

An example of a mapping type frame is shown below:

```
(*causative-inchoative
  (contain  *mapping-pattern1
            *mapping-pattern2))
```

The *causative-inchoative frame contains two mapping pattern frames, indicated by a *contain* link that includes *mapping-pattern1 and *mapping-pattern2.

3.4 Verb Class Frames

Verb class frames generalize over verbs with a common core sense and common syntactic behavior. Some example verb class frames (*verbs-of-breaking, *motion-path-verbs) are illustrated in Figure 1. The *verbs-of-breaking frame has an is-a link to the *causative-inchoative mapping type, indicating that verbs in the *verbs-of-breaking class can undergo the causative-inchoative alternation.

3.5 Lexical Frames

Lexical frames represent the language-dependent lexicon, and include pointers to corresponding conceptual frames. These frames also have is-a relations which link them to verb class frames, which are organized hierarchically according to the particular language.

The SEMANTICS slot in the lexical frame contains references to the conceptual frames associated with the lexical item. Particular restrictions on the meaning of the lexical item are captured by semantic role or feature assignment rules that may appear along with each SEMANTICS pointer.

For example, the SEMANTICS slot shown below for the verb *roll* points to the conceptual frame *MOVE. Included with the pointer to *MOVE is an assignment rule which indicates that the manner of *MOVE must have the meaning indicated by the conceptual frame *ROTATION. The *roll-1 frame has an is-a relation to the verb class frame, *motion-verbs.

```
(*roll-1
   (is-a *motion-verbs)
   (semantics
     (*MOVE
       (:manner = *ROTATION))))
```

More examples of lexical frames are shown in Figures 1. In Figure 1, *break-1 is a lexical frame, corresponding to the semantic notion *BREAK, which is a member of the *verbs-of-breaking verb class.

4 The Domain Conceptual Hierarchy

Conceptual frames represent knowledge of the world that is language-independent, for example, general concepts such as *EVENT and *PHYSICAL-OBJECT, as well as more specific concepts, like *BREAK and *SWIM[4]. Conceptual frames are organized hierarchically using inheritance relations. Selectional restrictions can be specified in conceptual frames, and appear as the fillers of semantic role slots.

5 Multiple Inheritance and Interpretive Mapping in Machine Translation

Our operational goals in constructing this hierarchy and its inheritance relations include the following:

[3] This is similar to the notion of *lexical forms* in lexical mapping theory (Bresnan and Kanerva, 1989), but the difference is that we incorporate case assignment rules into argument mapping rules to make the mapping a one step operation for use in generating or parsing sentences. In LFG, cases are assigned in each lexical entry through grammatical encoding theory, which identifies and assigns an appropriate case for a grammatical function in each lexical entry.

[4] An asterisk prefix is used to indicate frame names. Upper case frame names (e.g., *BREAK) indicate conceptual frames. Lower case is used for all other frame names (e.g., lexical frames, verb class frames, etc.).

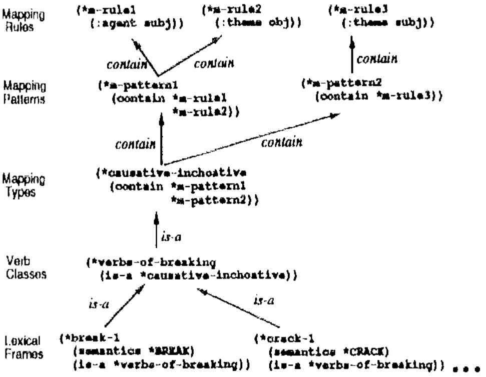
Mapping Rules
(*m-rule1 (:agent subj))
(*m-rule2 (:theme obj))
(*m-rule3 (:theme subj))
contain
contain
contain
Mapping Patterns
(*m-pattern1 (contain *m-rule1 *m-rule2))
(*m-pattern2 (contain *m-rule3))
contain
contain
Mapping Types
(*causative-inchoative (contain *m-pattern1 *m-pattern2))
is-a
Verb Classes
(*verbs-of-breaking (is-a *causative-inchoative))
is-a
is-a
Lexical Frames
(*break-1 (semantics *BREAK) (is-a *verbs-of-breaking))
(*crack-1 (semantics *CRACK) (is-a *verbs-of-breaking)) . . .

Figure 1: **Lexical Hierarchy Example: English**

- Support of rapid, straightforward acquisition of large amounts of lexical knowledge in an interactive environment;
- Elimination of unnecessary (and costly) redundancy in the representation of lexical knowledge.

5.1 Efficient Knowledge Acquisition

Productivity in the knowledge acquisition task is greatly enhanced by this hierarchical methodology. Rather than editing an ASCII file containing redundant mapping rule definitions for each lexical entry, the person entering new lexical concepts utilizes a 2-dimensional browsing and editing tool to add new knowledge to the system (Kaufmann, 1991).

Once the initial mapping rules, alternations and verb classes are specified, the user can easily link new lexical frames to existing verb classes, perhaps refining some of the knowledge in the upper portions of the hierarchy, but in general taking advantage of the compact nature of the hierarchy to avoid redundant data entry.

The frame representation presented here has a great advantage for the development of large-scale NLP systems, namely, that each mapping rule need only be defined once, and is thereafter inherited by all the lexical frames that require it. By positing intermediate levels of structure (mapping types and mapping patterns), significant generalizations can be captured which further enhance the compactness of the representation and the ease of knowledge acquisition.

5.2 Multiple Inheritance

The definition of containment, however, is not as straightforward as a simple *is-a* relation in traditional frame-based knowledge representation. The containment relation that obtains between mapping patterns and mapping rules is the usual conjunctive (multiple) type of inheritance, since a mapping pattern contains each and every mapping rule that it is linked to via a *contain* link. On the other hand, the containment relation that holds between mapping types and mapping patterns in *disjunctive*, since a mapping type contains different types of alternations, only one of which can be active at a given time for a particular verb. As a result, inheritance is performed in a different manner at these two levels in the hierarchy.

By default, FrameKit supports only conjunctive inheritance, which is most common in system where inheritance hierarchies are built using simple *is-a* links. We have developed user-defined inheritance methods for FrameKit that perform the appropriate inheritance operations at each level in the mapping hierarchy. When all of the possible subcategorization/mapping pairs must be retrieved for a given lexical frame, these inheritance methods perform the appropriate conjunctive inheritance, bundling the mapping rules together into mapping patterns, followed by disjunctive inheritance of mapping types to create any alternative readings of the lexical item. Simply speaking, the inheritance methods must recreate the explicit structure that is implicit in the inheritance hierarchy when it is necessary to represent distinct mappings for verbs at system run-time.

An example of how inheritance works at run time is illustrated in Figure 2. The two frames shown in the figure are instantiated by the inheritance methods from the lexical frame *break-1, and represent the two possible alternations of *break* (the causative reading and the inchoative reading).

```
(*BREAK-1153
   (THEME OBJ)
   (AGENT SUBJ)
   (SEMANTICS *BREAK)
   (CREATED-FROM *BREAK-1))
(*BREAK-1154
   (THEME SUBJ)
   (SEMANTICS *BREAK)
   (CREATED-FROM *BREAK-1))
```

Figure 2: **Instantiated Frames for *break-1**

5.3 Interpretive Mapping

The architecture in Figure 3 illustrates how our lexical hierarchy fits into the overall machine translation system. During parsing, the lexical entries stored in the source lexical hierarchy are accessed by the LFG parser; during the mapping of source f-structures to interlingua representations, the mapping rules in the lexical hierarchy are accessed by the mapper via instantiated mapping structures like those shown in Figure 2. During generation, the target language lexical hierarchy is utilized in a similar fashion. First, instantiated mapping structures are used to create target f-structures, and then target lexical entries are utilized by the LFG generator to produce target language strings.

6 Status

We have developed an extensive interpretive mapping hierarchy for Japanese, which includes 36 mapping rule frames,

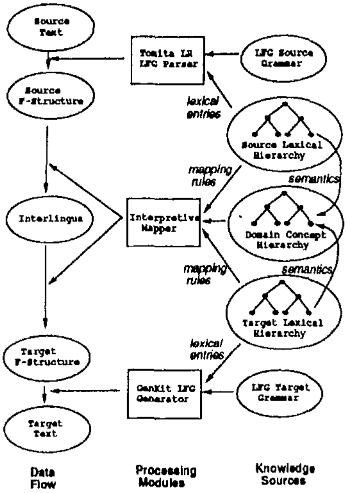

Figure 3: **System Architecture for Machine Translation**

45 mapping pattern frames, 37 mapping type frames, 54 verb class frames, and 100 lexical frames. Hundreds of additional lexical frames could be added to the hierarchy without modification of the existing hierarchical structure. We believe that our mapping frame hierarchy accounts for the syntactic behavior of a significant number of Japanese verb classes. The hierarchy is based on data for about 1000 verbs, taken from (Ishiwata and Ogino, 1983) and the IPAL report on basic Japanese verbs (IPAL, 1987).

We have also developed an initial mapping hierarchy for English verbs. The English and Japanese lexical hierarchies were utilized in the KBMT-89 system for the interpretation of Japanese sentences (Mitamura, et al., 1991). We are currently integrating our hierarchical structure into a large-scale system for translation of service manuals from English to Japanese. Since the argument mapping knowledge represented in our hierarchy is declarative rather than procedural, it can be used either in analysis or generation (cf. Figure 3).

7 Conclusion

High-quality machine translation requires an adequate semantic interpretation of the source text. To achieve this goal, we feel it is necessary to incorporate the kind of lexical knowledge and structure that has been explored in the theory of lexical semantics. We have presented a methodology that can be used to construct lexical hierarchies which represent lexical knowledge in a compact, efficient representation which captures relevant linguistic generalizations, as well as providing a useful framework for knowledge acquisition and system-building. This methodology is declarative and language-independent, and can be used either for parsing or generation.

References

[1] Bresnan, J. and J. M. Kanerva (1989) "Locative Inversion in Chichewa: A Case Study of Factorization in Grammar," *Linguistic Inquiry*, 20:1, 1-50.

[2] Carbonell, J. G. and M. Tomita (1987). "Knowledge-based Machine Translation: The CMU Approach," in Nirenburg, S. (ed.), *Machine Translation: Theoretical and Methodological Issues*, New York: Cambridge University Press.

[3] Fukui, N., S. Miyagawa, and C. Tenny (1985) "Verb Classes in English and Japanese: A Case Study in the Interaction of Syntax, Morphology and Semantics," *Lexicon Project Working Papers #3*, Center for Cognitive Science, MIT, Cambridge, MA.

[4] Goodman, K. and S. Nirenburg, eds. (1991) *A Case Study in Knowledge-Based Machine Translation*, San Mateo, CA: Morgan Kaufmann.

[5] Hale, K. and J. Keyser (1986) "Some Transitivity Alternations in English," *Lexicon Project Working Papers #7*, Center for Cognitive Science, MIT, Cambridge, MA.

[6] Ishiwata, T. and T. Ogino (1983) "Nihongo Yougen no Ketsugoka," *Bunpou to Imi I*, Asakura-shoten, Tokyo. 226-272.

[7] Jackendoff, R.S. (1983) *Semantics and Cognition*, MIT Press, Cambridge, MA.

[8] Jackendoff, R.S. (1987) "The Status of Thematic Relations in Linguistic Theory," *Linguistic Inquiry*, 18:3, 369-411.

[9] Johoushori shinkou jigyou kyoukai, ed. (1987) *Keisanki-you Nihongo Kihon-doushi Jisho IPAL - Basic Verbs: jisho-hen.*

[10] Kaufmann, T. (1991). *The ONTOS User's Guide.* Technical Memo, Center for Machine Translation, Carnegie Mellon University.

[11] Levin, B. (1985) "Lexical Semantics in Review: an Introduction," *Lexicon Project Working Papers #1*, Center for Cognitive Science, MIT, Cambridge, MA.

[12] Levin, B. (1987) "Approaches to Lexical Semantic Representation," Unpublished Manuscript.

[13] Levin, B. (1989) "English Verb Diathesis," *Lexicon Project Working Papers #32*, Center for Cognitive Science, MIT, Cambridge, MA.

[14] Mitamura, T. (1989) "The Hierarchical Organization of Predicate Frames for Interpretive Mapping in Natural Language Processing," Ph.D. dissertation, University of Pittsburgh.

[15] Mitamura, T., et al. (1991) "Analysis Lexicon," in Goodman and Nirenburg, eds., *A Case Study in Knowledge-Based Machine Translation*, San Mateo, CA: Morgan Kaufmann.

[16] Nyberg, E. (1988) *The FrameKit User's Guide: Version 2.0*, Technical Report, Center for Machine Translation, Carnegie Mellon University, CMU-CMT-88-MEMO.

[17] Rappaport, M. and B. Levin (1986) "What to Do with Theta-Roles," *Lexicon Project Working Papers #11*, Center for Cognitive Science, MIT, Cambridge, MA.

[18] Talmy, L. (1985) "Lexicalization Patterns: Semantic Structure in Lexical Forms," in T. Shopen, ed., Grammatical Categories and the Lexicon, *Language Typology and Syntactic Description 3*, Cambridge University Press.