[Human Language Technology, Plainsboro, 1994]

Building Japanese-English Dictionary based on Ontology for Machine Translation

Akitoshi Okumura, Eduard Hovy

USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

ABSTRACT

This paper describes a semi-automatic method for associating a Japanese lexicon with a semantic concept taxonomy called an ontology, using a Japanese-English bilingual dictionary as a "bridge". The ontology supports semantic processing in a knowledge-based machine translation system by providing a set of language-neutral symbols and semantic information. To put the ontology to practical use, lexical items of each language of interest must be linked to appropriate ontology items. The association of ontology items with lexical items of various languages is a process fraught with difficulty: since much of this work depends on the subjective decisions of human workers, large MT dictionaries tend to be subject to some dispersion and inconsistency. The problem we focus on here is how to associate concepts in the ontology with Japanese lexical entities by automatic methods, since it is too difficult to define adequately many concepts manually. We have designed three algorithms to associate a Japanese lexicon with the concepts of the ontology automatically: the equivalent-word match, the argument match, and the example match. We simulated these algorithms for 980 nouns, 860 verbs and 520 adjectives as preliminary experiments. The algorithms are found to be effective for more than 80% of the words.

1. Introduction

This paper describes a semi-automatic method for associating a Japanese lexicon with a semantic concept taxonomy using a Japanese-English bilingual dictionary as a "bridge", in order to support semantic processing in a knowledge-based machine translation (MT) system.

To enhance the semantic processing in MT systems, many system include conceptual networks called ontologies or semantic taxonomies [Bateman, 1990; Carlson and Nirenburg, 1990; Hovy and Knight, 1993; Klavans *et al.*, 1990; Klavans *et al.*, 1991; Knight, 1993]. These ontologies house the representation symbols used by the analyzer and generator. To put the ontologies to practical use, lexical items of each language of interest should be linked to appropriate ontology items. To support extensibility to new languages, the MT ontology should be language-neutral, if not language-independent[Hovy and Nirenburg, 1992]. However, the construction of language-neutral ontologies, and the association of ontology items with lexical items of various languages, are processes fraught with difficulty. Much of this work depends on the subjective decisions of more than one human workers. Therefore, large MT dictionaries tend to be subject to some dispersion and inconsistency. Many translation errors are due to these dictionary problems, because the quality of the MT dictionaries are essential for the translation process. If possible, the dictionary quality should be controlled by automatic algorithms during the process of development to suppress dispersions and inconsistencies, even if the final check should be entrusted to the human workers.

Another motivation for the development of automated dictionary/ontology alignment algorithms is the increased availability of online lexical and semantic resources, such as lexicons, taxonomies, dictionaries and thesa¡uri[Matsumoto *et al.*, 1993b; Miller, 1990; Lenat and Guha, 1990; Carlson and Nirenburg, 1990; Collins, 1971; IPAL, 1987]. Making the best use of such resources leads to higher quality translation with lower development cost[Hovy and Knight, 1993; Knight, 1994; Hovy and Nirenburg, 1992]. For example, the JUMAN system provides a Japanese unilingual lexicon for analyzing Japanese texts[Matsumoto *et al.*, 1993b]. The linkage of the unilingual lexicon to the ontology directly enables Japanese-English translation with lower development cost. From this viewpoint, automatic alignment algorithms represent a new paradigm for MT system building.

The problem we focus on here is how to associate concepts in the ontology with Japanese lexical entities by automatic methods, since it is too difficult to define adequately many concepts manually. We have designed three algorithms to associate a Japanese lexicon with the concepts of the ontology automatically: the equivalent-word match, the argument match, and the example match, by employing a Japanese-English bilingual dictionary as a "bridge". The algorithms make it possible to link the unilingual lexicons such as JUMAN with the ontology for the development of a Japanese-English MT system.

First, we describe three linguistic resources for developing the Japanese-English MT system: the ontology, the Japanese lexicon, and the bilingual dictionary. Next, we describe the automatic concept association algorithms for creating the MT dictionary. Finally, we report the results of the algorithms as well as future work.

2. Linguistic Resources

2.1. Ontology

At USC/ISI, we have been constructing an ontology, a large-scale conceptual network, for three main purposes with the PAngloss MT system, which we are building together with CMT and NMSU. The first is to define the interlingua constituents, which comprise the semantic meanings of the input sentences independent of the source and target languages. They are defined in the ontology as concepts that represent commonly encountered objects, entities, qualities, and relations. As the result of analyzing the input text, our MT system parsers produce interlingua representation using the concepts. The second purpose is to describe semantic constraints among concepts in the ontology, which works to support the analysis and generation processes of the MT system. The third purpose is to act as a common unifying framework among the lexical items of the various languages. The ontology is being semi-automatically constructed from the lexical database WordNet[Miller, 1990] and the Longman Dictionary of Contemporary English (LDOCE)[Knight, 1993]. At the current time, the ontology contains over 70,000 items. English lexical items are associated with over 98% of the ontology. The ontology is also being linked to a lexicon of Spanish words, using the Collins Spanish-English bilingual dictionary. In our work, it is being linked to the Japanese lexicon developed for the JUMAN word identification and morphology system[Matsumoto *et al.*, 1993b] by the algorithms described in this paper.

The ontology consists of three regions: the upper region (more abstract), the middle region, and the lower (domain specific) region. The upper region of the ontology is called the Ontology Base (OB) and contains approximately 400 items that represent generalizations essential for the various modules' linguistic processing during translation. The middle region of the ontology, approximately 50,000 items, provides a framework for a generic world model, containing items representing many English and other word senses. The lower regions of the ontology provide anchor points for different application domains. Both the middle and domain model regions of the ontology house the open-class terms of the MT interlingua. They also contain specific information used to screen unlikely semantic and anaphoric interpretations.

Japanese word	Bilingual Concept	English Word
jw_i	JW_i_001	ew_{11}, .., ew_{1p}
	JW_i_002	ew_{21}, .., ew_{2q}
	...	..
	JW_i_k	ew_{k1}, .., ew_{kr}
	...	..
	JW_i_n	ew_{n1}, .., ew_{ns}

Figure 1: Bilingual Word Correspondence

2.2. Japanese Lexicon

At USC/ISI, we employ the JUMAN morphological analyzer and the SAX parser for Japanese parsing[Matsumoto *et al.*, 1993b; Matsumoto *et al.*, 1993a]. These two modules use a lexicon of appropriate 100,000 Japanese words. The lexicon contains spelling/orthography forms, morphological information, and part-of-speech annotations. To be useful for MT, the Japanese words should contain English wordsense equivalents or semantic definitions. We provide this information required for linking JUMAN lexicon to the ontology concepts by employing a Japanese-English bilingual dictionary as a "bridge" .

2.3. Bilingual Dictionary

To link the unilingual Japanese JUMAN lexicon to the ontology, we employ a Japanese-English bilingual dictionary. This dictionary contains 75,000 words, providing Japanese-English word correspondences as shown in Figure 1. It is not difficult to link JUMAN lexical entries with the Japanese lexical items of the bilingual dictionary by a simple string matching. Our problem is: how can we automatically find the appropriate ontology item corresponding to each Japanese lexical item, if any ? Since we assume that there is at least one sense shared by a Japanese word jw_i and the equivalent English words, ew_{11}, ew_{12}, ... ew_{1j}, we define it as the bilingual concept JW_i_001. A bilingual concept JW_i_k is assigned to the kth correspondence pair. For each bilingual concept, we have extracted from the dictionary lists of the lexical information necessary for MT processing the Japanese word entry, including its definition, parts of speech, syntactic and semantic constraints for the arguments, English equivalent words including synonyms, and bilingual example sentences. The lexical lists indexed by the bilingual concept are shown in Figure 2.

For each bilingual concept, we replace information written in Japanese (such as the Japanese definition) by lists of English words for each Japanese word, by applying Japanese morphological analysis and the bilingual dictionary. Hereby we gain, for each Japanese word in the JUMAN lexicon that also appears in the bilingual dictio-

(Bilingual-concept TAMA_001
(Japanese-word "tama")
(Japanese-definition "a spherical object")
(Japanese-part-of-speech Noun)
(English-equivalent-words "a ball" "a globe")
(Examples "throw a ball" "catch a ball"
"hit a ball" "roll a ball"))

Figure 2: A bilingual concept for "Tama"

nary, the raw material to which we can apply algorithms to link it to the ontology.

3. Concept Association Algorithms

There are four cases on associating ontology concepts and equivalent bilingual concepts:

case-I Single to single association
A bilingual concept leads to one equivalent English word. The English word is linked to one ontology concept. Therefore, the bilingual concept is linked to one ontology concept as shown in Figure 3.

case-II Single to multiple association
A bilingual concept leads to one equivalent English word. The English word is linked to several ontology concepts. Therefore, the bilingual concept is linked to several ontology concepts as shown in Figure 4.

case-III Multiple to single association
A bilingual concept leads to several equivalent English words. The English words are linked to one ontology concept. Therefore, the bilingual concept is linked to one ontology concept as shown in Figure 5.

case-IV Multiple to multiple association
A bilingual concept leads to several equivalent English words. Each English word is linked to several ontology concepts. Therefore, the bilingual concept is linked to several ontology concepts as shown in Figure 6.

Case-I and case-III provide single associations between the bilingual concepts and the ontology concepts, which are simple. The problem is to associate the ontology concepts with equivalent bilingual concepts for case-II and

Bilingual Concept	English Word	Ontology concept
JW_i_k	ew_{k1}	EW_{k1}_0_1

Figure 3: Case-I: single to single association

Bilingual Concept	English Word	Ontology Concept
JW_i_k	ew_{k1}	EW_{k1}_0_1, .., EW_{kr}_0_t

Figure 4: Case-II: single to multiple association

case-IV. The equivalent-word match is designed for case-IV. The argument match and the example match are designed for case-II and for complementing the equivalent-word match.

3.1. Equivalent-word Match

The equivalent-word match algorithm is based on the algorithm developed by K.Knight for merging LDOCE and WordNet[Knight, 1993] and Knight's bilingual match algorithm[Knight, 1994]. The equivalent-word match searches for concept equivalencies by performing an intersection operation on all ontology concepts linked to the English equivalent words of the bilingual concept. Higher confidence is assigned to the concepts whose part of speech corresponds to the ontology type. For example, the Japanese noun "Tama" has nine senses in the dictionary. One of these senses is shown in Figure 7. The bilingual-concept TAMA_001 is represented by two English words: "ball" and "globe" . There are respectively six and three concepts for "ball" and "globe" in the ontology as shown in Figure 8. By intersecting the ontology concepts for a ball with the ontology concepts for a globe, TAMA_001 can be associated with the ontology concept ball_0_1 with a fairly high level of confidence.

3.2. Argument Match

The argument match collates Japanese argument constraints with ontology argument constraints. The argument match complements the equivalent-word match, because not all the lists contain two or more English equivalent words. For example, the Japanese verb "utsusu" has five senses in the dictionary. One of these senses is shown in Figure 9. There are three concepts linked to "infect" in the ontology as shown in Figure 10. Ontology concept infect_0_2 contains an argument constraint such as "Somebody infects somebody with

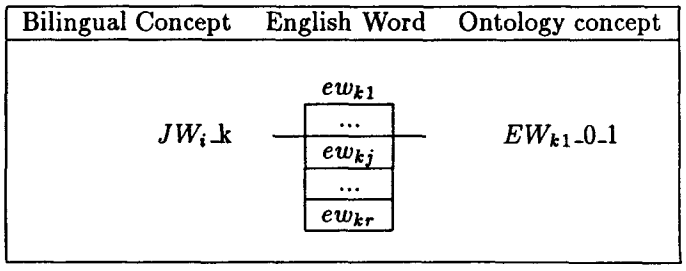

Figure 5: Case-III: multiple to single association

English word	Ontology Concept	Definition
ball	ball_0_1	round shape (a shape that is curved and without sharp angles)
	cotillion_0_1	cotillion (a lavish formal dance)
	clod_0_2	clod, glob, lump, chunk (a compact mass)
	ball_0_2	(a more or less rounded anatomical body or mass)
	ball_0_3	musket ball (a ball shot by a musket)
	ball_0_4	plaything, toy (an artifact designed to be played with)
globe	ball_0_1	round shape (a shape that is curved and without sharp angles)
	earth_0_4	earth, world (the planet on which we live)
	globe_0_1	(a sphere on which a map, esp. of the earth, is represented)

Figure 8: Ontology concepts and definitions for "ball" and "globe"

Bilingual Concept	English Word	Ontology Concept
JW_i_k	ew_{k1}	EW_{k1}_0_1, .., EW_{kr}_0_t
	..	..
	ew_{kj}	EW_{kj}_j-1_1, .., EW_{1p}_j-1_u
	..	..
	ew_{kr}	EW_{kr}_r-1_1, .., EW_{kr}_r-1_v

Figure 6: Case-IV: multiple to multiple association

some disease." When the algorithm matches the argument constraints, the ontology concept infect_0_2 is found to contain similar argument constraints to the bilingual concept UTSUSU_004. The algorithm assigns higher confidence to the association of UTSUSU_004 and infect_0_2.

3.3. Example Match

In order to complement the above two matches, the example match algorithm compares the bilingual examples with the ontology examples and definition sentences. By measuring the similarity of both examples, the algorithm determines the similarity of concepts. For example, the Japanese noun "ginkou" has one sense in the dictionary. The sense is shown in Figure 11. There are four concepts linked to "bank" in the ontology as shown in Figure 12. The algorithm calculates the similarity of two

```
(Bilingual-concept TAMA_001
    (Japanese-word "tama" )
    (Japanese-definition "a spherical object")
    (Japanese-part-of-speech Noun )
    (English-equivalent-words "a ball" "a globe")
    (Examples "throw a ball" "catch a ball"
              "hit a ball" "roll a ball"))
```

Figure 7: A bilingual concept for "Tama"

```
(Bilingual-concept UTSUSU_004
    (Japanese-word "utsusu" )
    (Japanese-part-of-speech Verb )
    (Japanese-constraints
        (Direct-Object Somebody)
        (Indirect-Object Disease))
    (English-equivalent-words "infect"))
```

Figure 9: One bilingual concept for "Utsusu"

word-sets (the words contained in the bilingual examples and the words contained in the ontology examples and definition sentence) by simply intersecting the two sets of words after transforming them to canonical dictionary entry forms and removing function words. In the case of GINKOU_001 example set and bank example sets, GINKOU_001 and bank_0_3 share the maximum number of words: "deposit" and "money". As a result, GINKOU_001 is highly associated with the ontology concept bank_0_3.

4. Results

We simulated these algorithms for 980 nouns, 860 verbs and 520 adjectives in a preliminary experiment. Half of the words belong to case-II and the other half to case-IV. The algorithms are applied according to the following procedure:

```
(Bilingual-concept GINKOU_001
    (Japanese-word "ginkou" )
    (Japanese-part-of-speech Noun )
    (English-equivalent-words "a bank")
    (Examples "deposit money in a bank"
              "have a bank account of 1,000,000 yen"
              "open an account with a bank"))
```

Figure 11: Bilingual concept for "Ginkou"

English word	Ontology Concept	Definition	Verb Frame
infect	infect_0_1	revolutionize, inspire, fill with revolutionary ideas	(SUB Somebody/Something) (DOBJ Somebody)
	infect_0_2	communicate a disease to	(SUB Somebody) (DOBJ Somebody) (with Disease)
	infect_0_3	taint, pollute	(SUB Somebody) (DOBJ Somebody)

Figure 10: Ontology concepts, definitions and verb frames for "infect"

1. The equivalent-word match is applied to case-II words. The results of the equivalent-word match are in Table 1.
2. The argument match is applied to all words except for the ones correctly determined by the equivalent-word match. The accuracy of the equivalent-word match and the argument match is in Table 2.
3. The example match is applied to all words except for the ones correctly determined by the above two matches. The total accuracy of the three matches is in Table 3.

Part of speech	Correct	Close	Open
Noun	51%	29%	20%
Verb	35%	38%	27%
Adjective	42%	33%	25%

Table 1: Accuracy by the equivalent-word match

- Correct: The highest confidence is assigned to all the correct concepts.
- Close: The highest confidence is assigned to some of the correct concepts.
- Open: No confidence value is assigned to the correct concepts.

Part of speech	Correct	Close	Open
Noun	51%	29%	20%
Verb	40%	38%	22%
Adjective	45%	33%	22%

Table 2: Accuracy after the argument match

Part of speech	Correct	Close	Open
Noun	55%	35%	10%
Verb	42%	38%	20%
Adjective	48%	37%	15%

Table 3: Total accuracy by the three matches

The algorithms are found to be effective for more than 80% of the words, thereby helping to reduce the dictionary development costs of human workers.

5. Discussion

In order to get better results, we are now improving the ratio of the open words and the close words from the following three viewpoints.

1. Semantic distance measurement
   To reduce the number of open words, the example match is being improved by using a more sophisticated algorithm for the semantic distance measured in the ontology[Resnik, 1993; Knight, 1993]. This measurement is also useful for improving the argument match, because the argument constraints are often described by the specific examples. In this case, the semantic distance measurement algorithm helps to determine whether the bilingual argument constraints are identical with the ontology argument constraints or not.
2. Other lexicons and databases
   For further improvement, other lexicons should be exploited. The open words usually are high ambiguity words with little information in the bilingual dictionary that have one equivalent English word with many meanings, with little constraint information and few examples. To compensate for the lack of information, we are now referring to other bilingual dictionaries and Japanese lexicons.
3. Integration of the three algorithms
   To reduce the number of close words, one integrated algorithm is being designed. By using the semantic distance measurement algorithm, one matching degree can be defined for both argument match and example match. Though the current equivalent-word match provides a high confidence only when all English-equivalent words share ontology concepts, we define the matching degree according to the number of English-equivalent words which can share ontology concepts. For example, when two of three English-equivalent words share an ontology concept EW_{kj}_1_1 and the other English-equivalent word is linked to an ontology concept EW_{kj}_2_1, a matching degree 0.66 is assigned to the association with EW_{kj}_1_1, and a matching degree 0.33 to EW_{kj}_2_1.

English word	Ontology Concept	Definition
bank	bank_0_1	(the sloping side of a declivity containing a large body of water)
	bank_0_2	(a long ridge or pile; "a bank of earth")
	bank_0_3	depository financial institution (a financial institution that accepts deposits and channels the money into lending activities)
	bank_0_4	array (an arrangement of aerials spaced to give desired directional characteristics)

Figure 12: Ontology concepts and definitions for "bank"

We determine the optimal weights for the three matching degrees based on the data used for simulation so that the integration algorithm can provide the most plausible association for the open words.

As well as improving these points, we are applying the algorithms to more words and other parts of speech. We plan to apply the algorithms to other bilingual dictionaries such as Chinese-English in order to increase the sophistication of the ontology for our multilingual MT system.

6. Acknowledgments

We would like to thank Kevin Knight for his significant assistance for this work. We also appreciate Kazunori Muraki of NEC Labs. for his support. This work was carried out under ARPA Order No.8073, contract MDA904-91-C-5224.

References

Bateman, J. 1990. Upper modeling: Organizing knowledge for natural language processing. In *Proc. Fifth International Workshop on Natural Language Generation, Pittsburgh, PA.*

Carlson, L. and S. Nirenburg. 1990. *World Modeling for NLP.* Tech. Rep. CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University.

Collins. 1971. *Collins Spanish-English/English-Spanish Dictionary.* William Collins Sons & Co. Ltd.

Hovy, E. and K. Knight. 1993. Motivating shared knowledge resources: An example from the pangloss collaboration. In *IJCAI-93 Workshop Large Knowledge Bases.*

Hovy, E. and S. Nirenburg. 1992. Aproximating an interlingua in a principled way. In *Proceedings of the DARPA Speech and Natural Language Workshop.* DARPA.

IPAL. 1987. *Lexicon of the Japanese Language for computers.* Information-technology Promotion Agency, Japan.

Klavans, Judith, Roy Byrd, Nina Wacholder, and Martin Chodorow. 1991. *Taxonomy and Polysemy.* Research Reportn RC 16443, IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598.

Klavans, Judith L., Martin S. Chodorow, and Nina Wacholder. 1990. From dictionary to knowledge base via taxonomy. In *Electronic Text Research.* Waterloo, Canada: University of Waterloo, Centre for the New OED and Text Research.

Knight, Kevin. 1993. Building a large ontology for machine translation. In *Proceedings of the ARPA Human Language Technology Workshop.* ARPA, Princeton, New Jersey.

Knight, Kevin. 1994. Merging linguistic resources. In Submitted to: *Proceedings of ACL'94* and *COLING'94.*

Lenat, D. and R.V. Guha. 1990. *Building Large Knowledge-Based Systems.* Reading, MA: Addison-Wesley.

Matsumoto, Y., Y. Den, and T. Utsuro. 1993. *Natural Language Parsing System SAX Manual, Ver.2.0.* Nagao Labs. Kyoto Univ. and Matsumoto Labs. AIST-Nara, Japan.

Matsumoto, Y., S. Kurohashi, T. Utsuro, H. Taeki, and M. Nagao. 1993. *Japanese Morphological Analysis System JUMAN Manual, Ver.1.0.* Nagao Labs. Kyoto Univ., Japan.

Miller, George. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4). (Special Issue).

Resnik, Philip. 1993. Semantic classes and syntactic ambiguity. In *Proceedings of the ARPA Human Language Technology Workshop.* ARPA, Princeton, New Jersey.