LEXICON-TO-ONTOLOGY CONCEPT ASSOCIATION USING A BILINGUAL DICTIONARY

Akitoshi Okumura[1]

Information Technology Research. Labs.
NEC Corp.
4-1-1 Miyazaki, Miyamae-ku
Kawasaki, Japan 216
email: okumura@hum.cl.nec.co.jp

Eduard Hovy

Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
email: hovy@isi.edu

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 with semantic information. To put the ontology to 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: the equivalent-word match, the argument match, and the example match.

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.

Many language processing systems include networks of concepts called ontologies or semantic taxonomies [Bateman 90, Carlson & Nirenburg 90, Hovy & Knight 93, Klavans et al. 90, Knight 93]. 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 must be linked to appropriate ontology items. To support extensibility to new languages, the MT ontology should be language-neutral, if not language-independent [Hovy & Nirenburg 92]. However, the construction of language-neutral ontologies, and the association of ontology items with lexical items of various languages, is difficult. Since much of this work depends on the subjective decisions of more than one human workers, large MT dictionaries tend to be subject to some dispersion and inconsistency.

[1] This work was done during a visit by the first author to the USC Information Sciences Institute.

Many translation errors are due to these dictionary problems. 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 is entrusted to humans.

Another motivation for automated dictionary/ontology alignment algorithms is the increased availability of online lexical and semantic resources, such as lexicons, taxonomies, dictionaries and thesauri [Matsumoto et al. 93b, Miller 90, Lenat & Guha 90, Carlson & Nirenburg 90, Collins 71, IPAL 87]. Starting with such resources leads to higher quality translation with lower development cost [Hovy & Knight 93, Knight 93]. For example, the JUMAN system provides a Japanese unilingual lexicon for analyzing Japanese texts [Matsumoto et al. 93b]. Linking JUMAN's lexicon to the ontology directly enables Japanese-English translation. From this viewpoint, automatic alignment algorithms represent a new paradigm for MT system building.

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, all three employing a bilingual Japanese-English dictionary as a "bridge". These 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

We have been constructing an ontology, a large-scale conceptual network, for three main purposes. The first is to define the interlingua constituents that comprise the semantic meanings of 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 90], the Longman Dictionary of Contemporary English (LDOCE) [Knight 93], and other resources. 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 bilingual Spanish-English dictionary. In our work, it is being linked to the Japanese lexicon developed for the JUMAN word identification and morphology system 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 and contains approximately 400 items that represent generalizations essential for the various

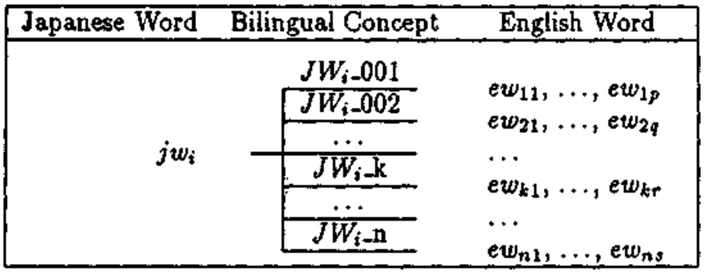

Figure 1: Bilingual Word Correspondences from Bilingual Dictionary

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.

2.2 Japanese Lexicon

We employ the JUMAN morphological analyzer [Matsumoto et al. 93b] and the associated SAX parser [Matsumoto et al. 93a] for Japanese parsing. These two modules use a lexicon of approximately 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 be linked to English wordsense equivalents or semantic definitions. We provide this information by employing a bilingual Japanese-English dictionary as a "bridge".

2.3 Bilingual Dictionary

To link the unilingual Japanese JUMAN lexicon to the ontology, we employ a bilingual Japanese-English 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 to automatically find the appropriate ontology item (if any) corresponding to each Japanese lexical item.

We proceed as follows. First we create a bilingual "concept" for each semantically distinct pairing of a Japanese word and its English equivalent(s). With this concept we then associate whatever information will help us find the equivalent ontology concept (if any). Finally we try to link the bilingual concept with its ontology equivalent.

In more detail, since we assume that there is at least one sense shared by a Japanese word jw_i and its dictionary equivalent English words ew_{11}, ew_{12}, ... ew_{1j}, we define this sense to be the bilingual "concept" JW_i_001. A bilingual concept JW_i_k is assigned to the *k*th correspondence pair. For each bilingual concept, we extract from the bilingual dictionary lists of appropriate lexical information, including its definition, parts of speech, syntactic and semantic constraints for the arguments, English equivalent words including synonyms, and bilingual example sentences. The

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

Figure 2: A Bilingual Concept for "Tama"

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 equivalent English words, by applying Japanese morphological analysis and the bilingual dictionary. This provides for each bilingual concept (and hence for each Japanese word in the JUMAN lexicon that also appears in the bilingual dictionary) the raw material to which we can apply our linking algorithms.

3 Concept Association Algorithms

There are four cases on associating ontology concepts with 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.

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.

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.

Cases I and III provide single associations between the bilingual concepts and the ontology concepts, which present no further difficulty. The problem is to associate the ontology concepts with equivalent bilingual concepts for cases II and 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.

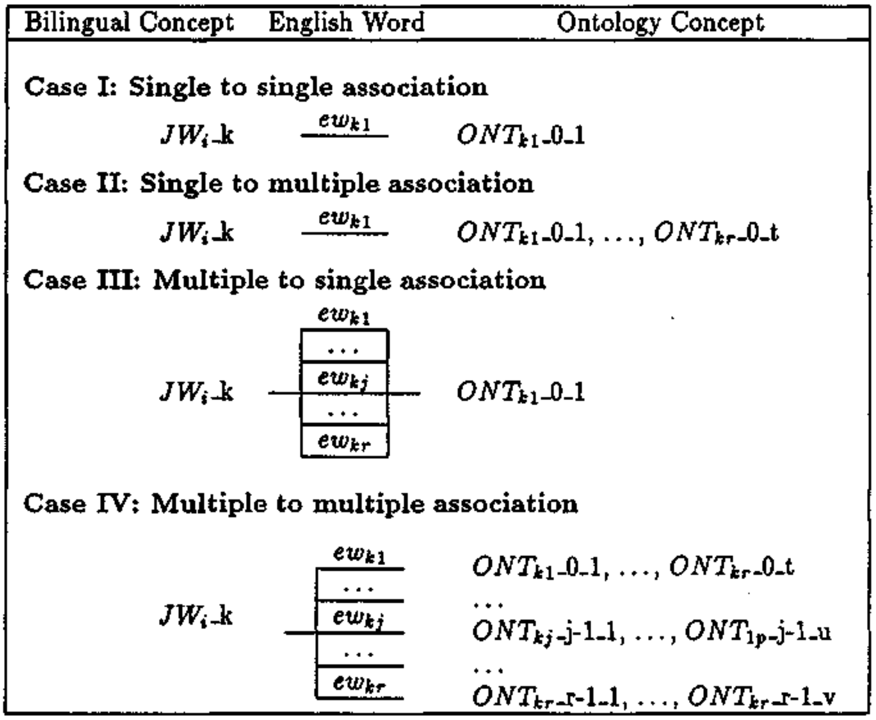

Figure 3: Four Cases of Word-to-Concept Association

3.1 Equivalent-word Match

The equivalent-word match algorithm is based on the algorithm developed by K. Knight for merging LDOCE and WordNet and Knight's bilingual match algorithm [Knight 93]. The equivalent-word match searches for concept equivalences 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 was shown in Figure 2. The dictionary lists for the bilingual-concept TAMA_001 two English words: "ball" and "globe". The Ontology has respectively six and three concepts for "ball" and "globe" (see Figure 4). By intersecting the ontology concepts for "ball" and "globe", TAMA_001 can be associated with the common ontology concept ball_0_1 with fairly high confidence.

3.2 Argument Match

The argument match compares Japanese argument constraints against 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 5. Ontology concept infect_0_2 contains an argument constraint such as "Somebody infects somebody with 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

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 4: Ontology Concepts and Definitions for "ball" and "globe"

confidence to the association of UTSUSU_004 and infect_0_2. The ontology contains three concepts linked to "infect", as shown in Figure 6.

```
(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 5: One bilingual concept for "Utsusu"

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

Figure 6: Ontology Concepts, Definitions and Verb Frames for "infect"

3.3 Example Match

In order to complement the above two matches, the example match algorithm compares the bilingual dictionary's example sentences with the ontology's examples and definition sentences. By measuring the similarity of 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 7. There are

```
(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 7: Bilingual Concept for "Ginkou"

English Word	Ontology Concept	Definition
bank	bank_0_1	(the sloping side of a declivity containing a 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 money into lending activities)
	bank_0_4	array (an arrangement of aerials spaced to give desired directional characteristics)

Figure 8: Ontology Concepts and Definitions for "bank"

four concepts linked to "bank" in the ontology as shown in Figure 8. The algorithm calculates the similarity of two 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 Current Work and Discussion

To estimate the potential benefits of these algorithms, we first performed a hand simulation with 980 verbs and 1380 verbs and adjectives. For this experiment we idealized the ontology, assuming enriched word differentiation, greater concept definition in some cases, and less differentiation in others. Although we realize that this simulation provides greatly overoptimistic results — around 50% correct and 35% close to correct (fewer than 10 alternatives, including the correct one) — it was valuable in identifying ways in which ontology shortcomings can be found through the dictionary.

Following this, after spending some time transforming the lexicon, dictionary, and several other similar resources to standard form so that they could be handled by the programs. M. Haines built a program that linked about 15,000 nouns of Cases I and II (using a combination of the Equivalent-word and Example matches) with encouraging results: 100% correct for 13.7% of the nouns (all of type Case I) and 13.8% (of type Case II), with decreasing confidence values for the remainder.

In order to improve results, we have considered the following three enhancements:

1. Semantic distance measurement: To reduce the number of open words, the example match can employ a more sophisticated measure for the semantic distance between concepts in the

ontology [Knight 93]. This measurement is also useful for improving the argument match, because the argument constraints are often described by the ontology concepts as well.

2. Other lexicons and databases: Using additional definitions and knowledge from other resources will tend to reduce the ambiguity of open-class words with one English equivalent with many senses.

3. Integration of the three algorithms: Since the algorithms use different kinds of knowledge, they can fruitfully be combined. In some cases, it is useful to run one algorithm to some level of quality, use its results to partition the data for another algorithm and then to take the results of the second back to the first for further processing, and so on. We have not yet studied the full implications of such alternation.

5 Acknowledgments

We would like to thank Kevin Knight and Matthew Haines for their significant assistance with this work. We also appreciate Kazunori Muraki of NEC Labs. for his support.

References

[Bateman 90] Bateman, J. 1990. Upper modeling: Organizing knowledge for natural language processing. In *Proc. Fifth International Workshop on Natural Language Generation, Pittsburgh, PA.*

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

[Collins 71] Collins. 1971. *Collins Spanish-English/English-Spanish Dictionary.* William Collins Sons & Co.

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

[Hovy & Nirenburg 92] Hovy, E. and S. Nirenburg. 1992. Approximating an interlingua in a principled way. In *Proceedings of the DARPA Speech and Natural Language Workshop.* DARPA.

[IPAL 87] IPAL. 1987. *Lexicon of the Japanese Language for computers.* Information-technology Promotion Agency, Japan.

[Klavans et al. 90] 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 93] Knight, Kevin. 1993. Building a large ontology for machine translation. In *Proceedings of the ARPA Human Language Technology Workshop.* ARPA, Princeton, New Jersey.

[Lenat & Guha 90] Lenat, D. and R.V. Guha. 1990. *Building Large Knowledge-Based Systems.* Reading, MA: Addison-Wesley.

[Matsumoto et al. 93a] 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 et al. 93b] 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 90] Miller, George. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4) (Special Issue).