A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons

Yoshihiko Hayashi[§¶], Toru Ishida[*¶]

[§]Graduate School of Language and Culture, Osaka University
1-8 Machikaneyama-cho, Toyonaka-shi, Osaka 560-0043, Japan
hayashi@lang.osaka-u.ac.jp

*Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
ishida@i.kyoto-u.ac.jp

[¶]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289, Japan

Abstract

The *Language Grid*, recently proposed by one of the authors, is a language infrastructure available on the Internet. It aims to resolve the problems of accessibility and usability inherent in the currently available language services. The infrastructure will accommodate an operational environment in which a user and/or a software agent can develop a language service that is tailored to specific requirements derived from the various situations of intercultural communication. In order to effectively operate the infrastructure, each atomic language service has to be discovered by the planner of a composite service and incorporated into the composite service scenario. Meta-description of an atomic service is crucial to accomplish the planning process. This paper focuses on dictionary access services and proposes an abstract dictionary model that is vital for the accurate meta-description of such a service. In principle, the proposed model is based on the organization compatible with Princeton WordNet. Computational lexicons, including the EDR dictionary, as well as a range of human monolingual/bilingual dictionaries are uniformly organized into a WordNet-like lexical concept system. A modeling example with a few dictionary instances demonstrates the fundamental validity of the model.

1. Introduction

Several language services are being made available on the Internet. Some of the representative types of language services include text translation and dictionary access. To help reduce language barriers that prevent smooth and effective communication, it is only natural to adopt these services in various situations of intercultural collaboration. The reality, however, proves to be different; it is substantially difficult for a user to find a language service that precisely meets his/her requirements that emerge from communicative situations. Furthermore, it is usually impossible for a user to newly define or create such a language service. There exist problems of accessibility and usability (Ishida, 2006).

The *Language Grid* recently proposed by Ishida (2006) is essentially a language infrastructure available on the Internet. It aims to resolve the problems by accommodating an operational environment in which a user, as well as a software agent, can develop a language service that is tailored to specific requirements by developing appropriate atomic language services and coordinating them properly.

In order to achieve this goal, we have to solve a variety of problems, ranging from a service protocol that was optimized for efficient language service execution to higher level management functions that deal with intellectual property rights and/or pricing policies. However, one of the most essential issues among them is the development of a planning mechanism that will enable a user/agent to define a composite language service effectively. With the planning mechanism, a user/agent will be able to create a plan or scenario that specifies the execution of appropriate atomic services in a proper arrangement, given a situation in which each atomic service is adequately described. This description, known as metadata, should be provided based on an ontology that ensures a common understanding between users and agents (Buietlaar, 2003).

This paper concentrates on dictionary access services, and proposes a unified abstract dictionary model that can represent a range of machine-readable human dictionaries and computational concept lexicons[1]. The model will provide us with a strong foundation on which we can accurately meta-describe dictionary/lexicon language resources that are the essential components of any dictionary service. In addition, semantic wrappers for dictionary services could be efficiently generated by referring to model-based meta-descriptions.

The remainder of this paper is organized as follows. Section 2 outlines the language grid that motivates the reported work, and discusses the fact that dictionary services are in high demand in several situations of intercultural collaboration. Section 3 describes in detail the types and instances of the language resources that are being considered in the language grid. Section 4 proposes an abstract dictionary model and presents a modeling example. Section 5 discusses several research issues while referring to related works. Finally, concluding remarks are presented in section 6.

1 We distinguish between the terms “lexicon” and “dictionary” when necessary: a lexicon implies a set of formalized entries that are used by computer programs, whereas a dictionary refers to a physical or data object that provides lexical information to human users. This notion is almost compatible with the one introduced by Wilks (1996: 6).

2. Language Grid and Dictionary Services

2.1. Overview of the Language Grid

Figure 1 exemplifies an imaginary organization of the language grid in terms of its architecture (Ishida, 2006). As shown in the figure, the language grid has two dimensions: "horizontal" and "vertical." The horizontal dimension is associated with repertoires of languages, whereas the vertical dimension represents its application fields/purposes. Ishida (2006) designates the former as "standard languages" and the latter as "community languages."

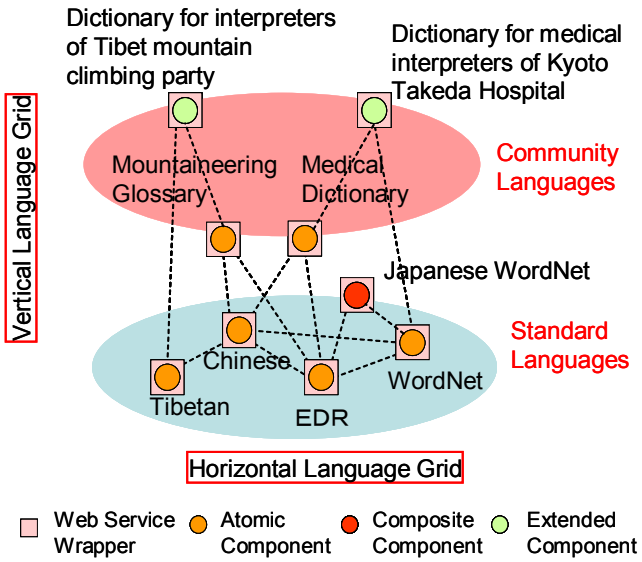

Figure 1: Language Grid Architecture (Ishida, 2006)

As implied in Figure 1, the key to the success of the language grid scenario is the extent to which we can effectively define a composite component. An example of a composite service shown in Figure 1 is "Japanese WordNet"; a dictionary service with which users can access a Japanese concept lexicon (EDR, 2003) like a Japanese equivalent of the English WordNet (Fellbaum, 1998). It is possible to achieve this by coordinating an EDR dictionary access service with a WordNet consultation service, probably with the help of a Japanese-to-English bilingual dictionary service.

2.2. Dictionary Services in the Language Grid

Ishida (2006) describes some pilot studies that involve a few NPO-based communities that are planned such that requirements can be derived from the communities to the language grid. The application fields of the studies range from medical interpretation, children's communication, and text-based communication in more than ten languages. While machine translation is the most required language service, dictionary access services could be more in demand. Off-the-shelf Machine Translation systems are still limited to popular language pairs, and often the MT output is insufficient, in particular, for rigorous communications such as those observed in medical interactions. In short, access to bilingual dictionaries is essential in the language grid. In addition, access to concept and/or monolingual dictionaries will be of great help to users to consult the meaning and usage of linguistic expressions. An appropriate combination of these dictionaries could be highly useful for the language grid users facing language barriers.

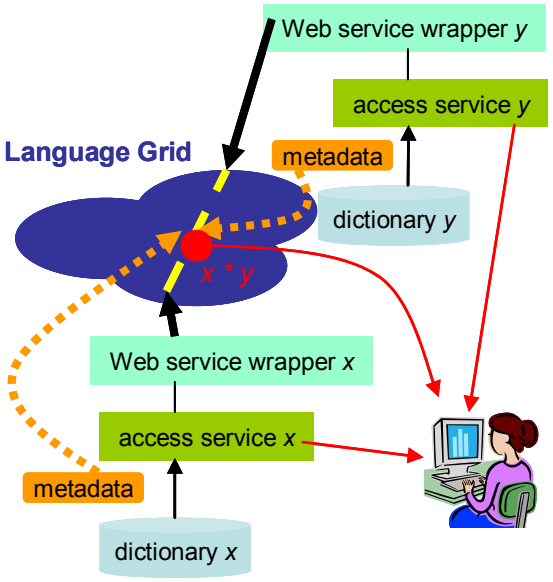

Figure 2: General Structure of Dictionary Services

2.3. Metadata for a Dictionary Service

Figure 2 schematizes a general structure of dictionary services in the language grid. The user can access atomic dictionary services that were offered by the service providers. He/she can also consult a composite dictionary service that is made available by the language grid.

The metadata attached to an atomic service plays a central role in this configuration; the metadata is retrieved and utilized whenever a composite service is to be defined. It should not only describe the service grounding information (The OWL Service Coalition, 2003) but also disseminate semantic features of the associated language resources. The dictionary model proposed in this paper provides a strong foundation with which the semantic features can be properly encoded. In addition, this configuration may lead to the possibility of highly effective service deployment; the Web service wrapper of a dictionary service could be efficiently generated by a computational process that examines the metadata.

3. Dictionary/Lexicon Resources

As mentioned in 2.2, it will be effective for language grid users to be able to combine dictionaries primarily compiled for human use along with computational dictionaries, each classifying the concepts in a language.

3.1. Machine-Readable Dictionaries

Currently, there are numerous access services to the machine-readable version of a printed dictionary (MRD; Machine-Readable Dictionary) are available on the Web. The following examples are taken from the "goo dictionary service"[2] that is based on Sanseido[3] dictionaries.

2 http://dictionary.goo.ne.jp/

Figure 3 introduces an entry for the English word form "bank" as a noun in an English-to-Japanese bilingual dictionary[4]. It shows that there are five senses of the word, and each sense is separated by a semicolon. The word sense "financial institution" is listed in the first item.

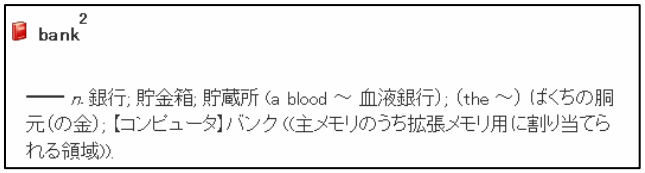

Figure 3: An Entry from an English-to-Japanese Bilingual Dictionary Service

Figure 4 shows an entry for the same word form in an English dictionary service[5]. In this dictionary, there are five sense groups (one is marked as obsolete) or nine fine-grained senses[6].

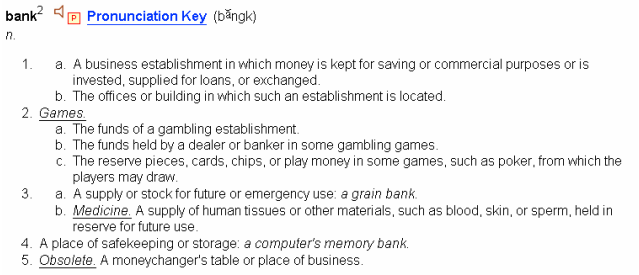

Figure 4: An Entry from an English Dictionary Service

As typically shown in these examples, the information structure in an MRD entry is usually organized based on word senses, with more or less variations in the sense discrimination criteria that are dictionary dependent.

3.2. Computational Concept Dictionary

Most of the concept dictionaries have been compiled to be used by computer programs; hence, they could be termed as CCL (Computational Concept Lexicons). We focus on Princeton WordNet and EDR electronic dictionaries as representative and useful CCL resources.

WordNet

WordNet is a well-known and the most utilized lexical concept system for English. A lexical concept, represented as a node in the system, is defined by a set of word forms termed *synset*. The existence of a word in a synset implies that one of the word senses of the word form is associated with the lexical concept. Synsets are linked by various conceptual relations, such as hyponymy-hypernymy, meronymy-holonymy, and antonymy. A lexical node is annotated by a natural language description called *gloss* that substantially helps a user to understand the concept.

The EDR Electronic Dictionary

The EDR electronic dictionary is a dictionary system developed for NLP systems. The entire dictionary system consists of the following five large-scale dictionaries (or dictionary groups): Word (monolingual), Bilingual, Concept, Co-occurrence, and Technical Terminology dictionaries. The dictionaries are in Japanese and English. One of the most prominent features of the dictionary system is that every entry is associated with a unique *concept identifier* (CID), assuming that there is a unique concept system that covers both Japanese and English well. The Concept Classification dictionary, that is main component of the Concept dictionary, organizes the entire concept system into a taxonomy in which a node represents a concept, and a link connects the concepts that have a super-sub relation.

A concept node, if it can be defined, is labeled *headconcept*. Every entry in any monolingual and bilingual dictionaries in the EDR system is indexed by a *headword* and has a CID that links the headword to a concept node in the concept system. This enables us to form a pseudo-synset for a concept node. The headconcept together with the headwords in the EDR dictionaries can constitute a set of synonyms[7].

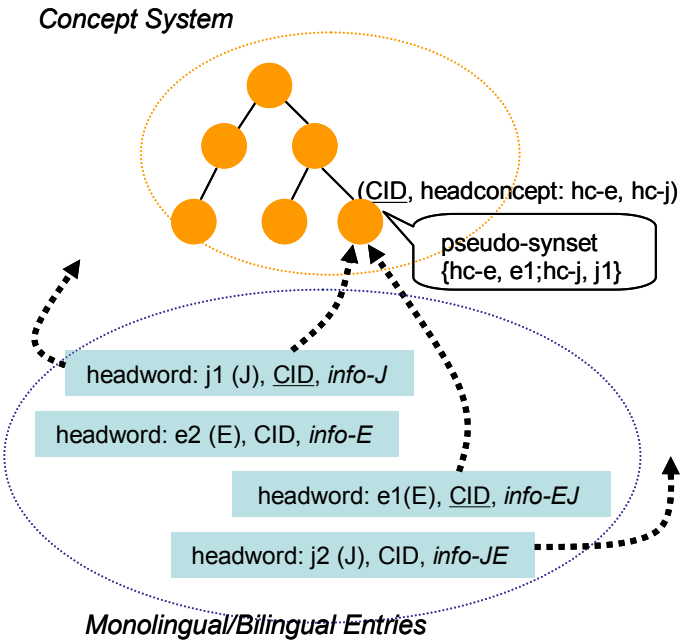

Figure 5: Logical Structure of the EDR Dictionary

Figure 5 depicts the logical structure of the EDR dictionary system. In Figure 5, there are two entries that share an identical CID: One entry is from the Japanese monolingual dictionary that has "j1" as its headword, and the other is from the English-to-Japanese bilingual dictionary that has "e1" as the headword. The concept node with this CID has an English headconcept "hc-e" as well as a Japanese headconcept "hc-j."

3 Sanseido Co., Ltd, http://www.sanseido-publ.co.jp/

4 The word form "bank" in the "sloping land" sense is described in a separate entry which indicates that it is a homonym, rather than in the entry for one sense of the word form.

5 http://dictionary.reference.com/

6 Nest levels are currently ignored; only the deepest (leaf level) items are considered as word senses in the model.

7 Unlike WordNet, a concept node with an empty synset can be possible in the EDR dictionary system. In fact, over 8,000 nodes (approximately 2% of all nodes) have empty synsets.

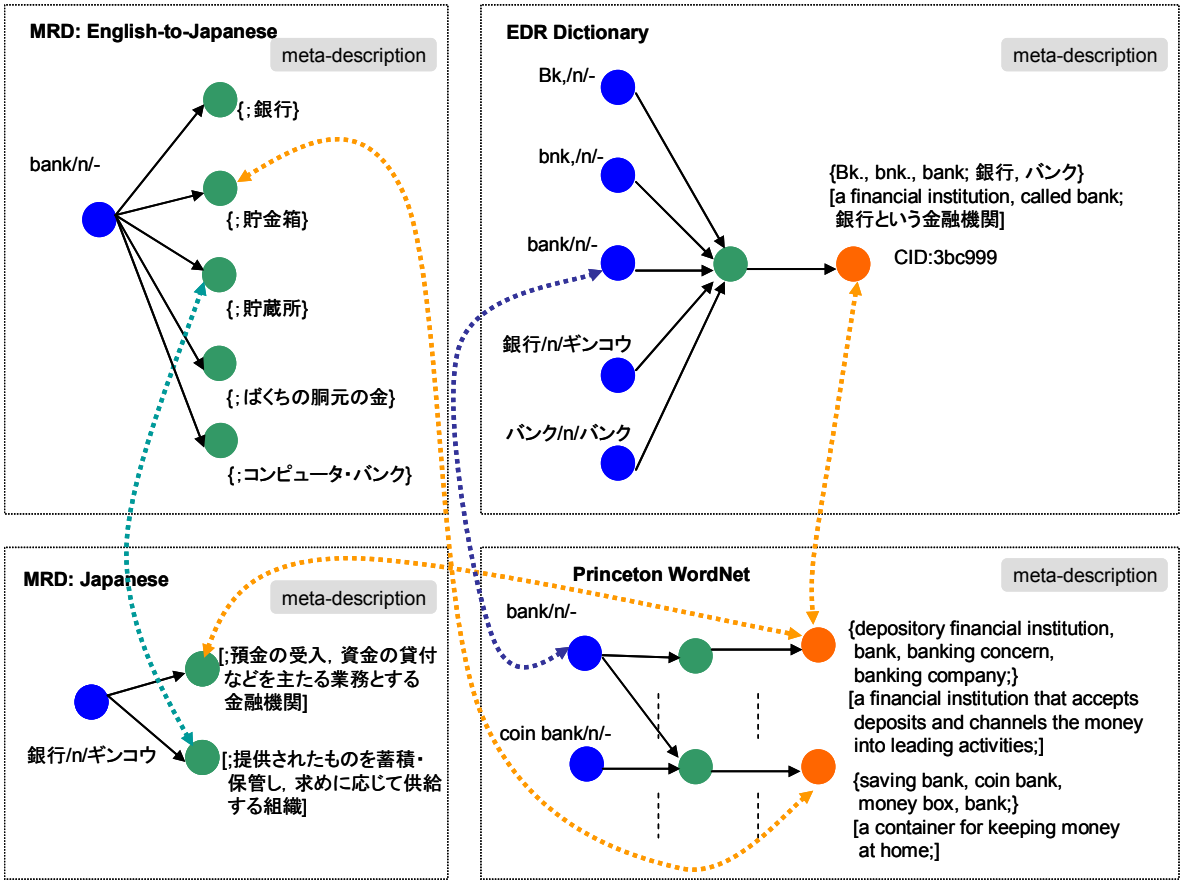

Figure 6: An Example of Dictionary Modeling

These headwords and headconcepts form a pseudo-synset "{hc-e, e1, hc-j, j1}" that represent the concept identified by the CID. This suggests that the logical structure of the EDR dictionary is almost compatible with the WordNet basic structure, implying that our dictionary model can be and should be based on the WordNet organization. It should be noted that the pseudo-synset is a language mixture of Japanese and English.

4. The Dictionary Model

This section presents the basic ideas of our dictionary model that represents the dictionary/lexicon instances covered by the language grid interests.

4.1. Overview of the Model

A similar lexicon model was recently proposed is the LMF (Lexical Markup Framework) (Francopoulo, 2006; ISO 24613, 2005). Our model has the three fundamental principles of the LMF: simplicity/clarity, universal expressiveness, and scalability. However, the LMF is described as "an abstract metamodel that provides a common, standardized framework for computational lexicons." This is in contrast with our model; our primary concern is to represent semantic/conceptual information that is useful for users of MRDs as well as for computational lexicons. Further, we would also like to represent the derived relations that are a result of computational processes that attempt to relate dictionary entries across different dictionaries. These derived relations can be accumulated within the language grid environment for future use. Therefore, the model should be able to encode these relations.

Figure 6 introduces an example of dictionary modeling and also suitably provides an overview of our dictionary model. Figure 6 presents entries from four dictionaries/lexicons, which are in some way related to the English word "bank" in the form of a graph structure. The entire model space is divided into individual dictionary spaces. Each space represents an instance from a dictionary/lexicon and is summarized by a dictionary meta-description (as shown in gray boxes). The meta-description has not been provided in detail in this paper; however, it should include information such as dictionary ID (probably given in a URI (Uniform Resource Identifier)), type (monolingual, bilingual, concept, etc.), application domain, language(s), character encoding scheme, and inventory of lexical/semantic/conceptual relations. In addition, administrative information about the resource should also be included here.

4.2. Nodes

Nodes in a dictionary space are classified as a "lemma node," "sense node," or "concept node."

- Lemma node (the blue node in Figure 6): A lemma node is usually associated with the headword of a dictionary entry. It is identified by a canonical written form and part of speech, if defined. In

addition, words in Japanese (or any other language similar to Japanese) are further defined by "katakana reading": reading of the written canonical form in katakana. This is necessary for a Japanese written form that has more than one pronunciation and associated meaning. In Figure 5, a lemma node is marked by a label with the form of "/canonical-form/part-of-speech/katakana-reading." A lemma node has one or more links to sense nodes since an entry in a dictionary can have several sub-entries, with each representing a distinct word sense.

- Sense node (the green node in Figure 6): A sense node stores most of the content described in the MRD. In other words, if it is from a bilingual dictionary, it stores gloss, usage examples, collocations, and translations. For a CCL like WordNet or EDR, a sense node is just an intermediate node that simply connects a lemma node to a concept node.
- Concept node (the orange node in Figure 6): The information associated with a lexical concept is encapsulated in a concept node. It stores gloss, usage examples, and semantic/conceptual relations, as described in the lexicon. The synset for a concept node is represented by the set of incoming links from the sense nodes.

In Figure 6, a sense/concept node is labeled by a synset and a gloss. The former is represented by "{ }," and the latter by "[ ]"; within the parentheses, the English and Japanese parts are separated by semicolons.

At present, we concentrate on semantic/conceptual information that may be useful in cross-language/intercultural communications. However, the following are the types of information that might further facilitate communications: information on etymology, phonology, morphology, syntax, and syntax-semantics linking. For instance, etymological information might be useful in explaining a highly culturally dependent concept that foreigners find difficult to understand. Further, phonological information, such as pronunciations, could be provided by recorded/synthesized voices in order to directly assist non-native users to speak. Information on syntax-semantics linking may provide users with fine-grained knowledge to compose a fluent phrase.

4.3. Links

The fundamental relations between nodes in an initial dictionary model will be lemma-to-sense for most of the MRDs and sense-to-concept and concept-to-concept for the CCLs. These relations are identified by dictionary entry parsing or simple reformatting of the lexicon source data. Sometimes relations like antonyms or synonyms are explicitly described in a dictionary by using special symbols (e.g., ⇔) that link two sense nodes.

As mentioned previously, some relations that cross dictionary spaces are identified as a result of a composite language service execution and can be attached to the entire dictionary space. These derived relations can be accumulated in the language grid environment and may be used in the future. In Figure 6, these relations are indicated by the dotted arrows. The following are the types of derived links:

- Lemma-to-lemma link: Equivalent relations of this type are monolingual and are identified by a string level matching process. Grammatical derivational relations can also be encoded with this type of link.
- Sense-to-sense link: Equivalent or near-equivalent relations of this type can either be monolingual or cross-lingual. In order to provide a composite language service like "Japanese-to-Arabic dictionary" by cascading a Japanese-to-English dictionary and an English-to-Arabic dictionary, we need to identify the optimal sense-to-sense relation between these dictionary entries and represent it with this type of link.
- Sense/concept-to-concept link: This type of relation is also monolingual or cross-lingual. However, cross-lingual relations are more important in our context. Let us assume that we have to provide a composite language service like "Japanese WordNet" that looks for the most similar concept in Princeton WordNet when provided with a Japanese word. It is likely that we will first search for possible English senses in Japanese-to-English and/or English dictionaries and then align them with candidate concept nodes in the WordNet space. The discovered relations should be represented by this type of link.

Since most of the derived relations are identified by computational processes rather than by experts, we require a description for the derived link that indicates the manner in which the relation is reliable. This can be done by assigning weight values that reflect confidence and/or quality assessment of the process to the derived relational links. Another important issue is the inventory of the derived relations that will be defined probably by consulting the mapping relations in the EuroWordNet (Vossen, 2004).

5. Discussions

Several activities associated with the standardization of computational lexicons, such as MILE (Multilingual ISLE Lexical Entry) (Calzolari, 2002) and OLIF (Open Lexicon Interchange Format) (McCormick, 2005) have been reported. Among them, the LMF (Francopoulo, 2006; ISO 24613, 2005) is the most related work that is currently being pursued in ISO TC 37/SC 4. It adopts a two-layered approach; a "form" node provides access to information associated with a word form, whereas a "sense" node encodes its semantic/conceptual information. On the other hand, our model adopts a three-layered approach. A "lemma" node functions in almost the same manner as that of the "form node" in the LMF. In our model, a "sense" node is introduced to represent an MRD entry that is given according to the word senses, while a "concept" is employed to encode a lexical concept node in the CCLs. However, this difference is not innate; thus, a sense node in our model can be diverted into a concept node because any sense-to-concept relations are based on a one-to-one relation.

One of the problematic issues in dictionary modeling, particularly with cross-linguality, may be the problem of the "lexical gap" (Janssen, 2004). For example, the fourth sense of "bank" in the English-to-Japanese dictionary shown in Figure 6 has no direct translations in Japanese.

The translation provided in the entry is somewhat explanatory, roughly implying the "money of a gambling dealer." Thus, the sense node in the English-to-Japanese dictionary may not be able to find its corresponding node in any Japanese dictionary spaces. This sense is defined in WordNet (Ver.2.1) as the eighth sense of "bank," and explained as "the funds held by a gambling house or the dealer in some gambling games."

However, given the scope of the language grid, this might not be a serious problem. The language grid is intended for use in intercultural communications. The explanatory translation may make complete sense to a Japanese user whenever he/she is on the comprehension side of the communication. In contrast, when he/she is on the production side, he/she may not necessarily choose the word "bank" to describe the sense of the "funds held by a gambling dealer" sense. Exact words such as "money of a gambling dealer" would be sufficient, even though it may not be very felicitous. Nevertheless, the lexical gap is certainly an important and interesting linguistic issue that emerges while detailing the dictionary model. We will eventually have to develop a way to deal with multi-word expressions that are used in explanatory translations.

6. Concluding Remarks

This paper proposed an abstract dictionary model that can represent machine-readable human dictionaries, such as monolingual and bilingual dictionaries, as well as computational concept dictionaries, such as Princeton WordNet or the EDR electronic dictionary. In principle, the model is based on the organization that is compatible with WordNet, insisting that the EDR dictionaries can also be reorganized into a WordNet-like lexical concept system. A modeling example with four dictionary instances was provided to demonstrate the fundamental validity of the model.

However, the proposed model is still in its early stage, and only a rough sketch of the model has been introduced in this paper. We will have to extend it to cover a wide range of dictionary/lexicon instances. We may have to refine it linguistically, particularly to deal with the lexical gaps. In the course of our study, we should be aware and should possibly cooperate with the ongoing standardization activities such as the LMF.

At the same time, we also have to accomplish several tasks to achieve the goals of the language grid. First, we should formalize the model and implement actual dictionary services by using the frameworks developed in the scope of the Semantic Web service. The work reported to represent WordNet with RDF/S and OWL (WordNet Task Force, 2004) will be helpful in achieving this goal. Second, we should apply computational processes to link lexical information in different dictionaries based mainly on word sense/lexical concepts. Studies that have achieved this task include Utiyama (1997), Chen (2002) and others. Further, to facilitate the deployment of dictionary access services in the language grid, we should provide a set of tools for efficiently constructing Web service wrappers. In an attempt to solve this problem, we may be able to adopt the ideas of "programming by demonstration" (Bauer, 2000) and "scheme-guided wrapper generation." (Meng, 2002)

References

Bauer, M., Degler, D., Paul, G., and Meyer, M. (2000). Programming by Demonstration for Information Agents. *Communications of the ACM* Vol.43, No.3, pp.98–103.

Buitelaar, P., Declerck, T., Calzolari, N., and Lenci, A. (2003). Language Resources and the Semantic Web. In *Proceedings of ELSNET/ENABLER workshop.*

Calzolari, N., Zampolli, A., and Lenci, A. (2002). Towards a Standard for Multilingual Lexical Entry: The EAGLES/ISLE Initiative. In *Proceedings of CICLing 2002* pp.264–279.

Chen, H., Ling, C., and Lin, W. (2002). Building a Chinese-English WordNet for Translingual Applications. *ACM Transactions on Asian Language Information Processing* Vol.1, No.2, pp.103–122.

EDR (2003). EDR Electronic Dictionary Technical Guide. http://www2.nict.go.jp/kk/e416/EDR/.

Fellbaum, C. (Eds.) (1998). WordNet: An Electronic Lexical Database. MIT Press.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).*

Ishida, T. (2006). Language Grid: An Infrastructure for Intercultural Collaboration. In *Proceedings of IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)* keynote address, pp.96–100.

ISO 24613. (2005). Lexical Resource Management–Lexical Markup Framework (LMF). Working document: ISO/TC 37/SC 4 N130 Rev.7.

Janssen, M. (2004). Multilingual Lexical Databases, Lexical Gap, and SIM*u*LLDA. *International Journal of Lexicography* Vol.17, No.2, pp.137–154.

McCormick, S. (2005). The Structure and Content of the Body of an OLIF v.2.0/2.1 File. OLIF Consortium. http://www.olif.net/olif2.1tmp/documentation.htm.

Meng, X., Lu, H., Wang, H., and Gu, M. (2002). SG-WRAP: A Schema-Guided Wrapper Generator. In *Proceedings of the 18th International Conference on Data Engineering (ICDE'02)* pp.331–332.

The OWL Services Coalition. (2003). OWL-S: Semantic Markup for Web Services. http://www.daml.org/services/owl-s/1.0/owl-s.html.

Utiyama, K., and Hasida, K. (1997). Bottom-up Alignment of Ontologies. *In Proceedings of IJCAI-97 Workshop on Ontologies and Multilingual NLP*, pp.35–40.

Vossen P. (2004). EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography* Vol.17, No.2, pp.161–173.

Wilks, Y., Slator, B.M., and Guthrie, L.M. (1996). Electric Words: Dictionaries, Computers, and Meanings. MIT Press.

Wordnet Task Force. (2004). Wordnet in RDFS and OWL. W3C Editor's Draft. http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.htm