Infrastructure for standardization of Asian language resources

Tokunaga Takenobu
Tokyo Inst. of Tech.

Virach Sornlertlamvanich
TCL, NICT

Thatsanee Charoenporn
TCL, NICT

Nicoletta Calzolari
ILC/CNR

Monica Monachini
ILC/CNR

Claudia Soria
ILC/CNR

Chu-Ren Huang
Academia Sinica

Xia YingJu
Fujitsu R&D Center

Yu Hao
Fujitsu R&D Center

Laurent Prevot
Academia Sinica

Shirai Kiyoaki
JAIST

Abstract

As an area of great linguistic and cultural diversity, Asian language resources have received much less attention than their western counterparts. Creating a common standard for Asian language resources that is compatible with an international standard has at least three strong advantages: to increase the competitive edge of Asian countries, to bring Asian countries to closer to their western counterparts, and to bring more cohesion among Asian countries. To achieve this goal, we have launched a two year project to create a common standard for Asian language resources. The project is comprised of four research items, (1) building a description framework of lexical entries, (2) building sample lexicons, (3) building an upper-layer ontology and (4) evaluating the proposed framework through an application. This paper outlines the project in terms of its aim and approach.

1 Introduction

There is a long history of creating a standard for western language resources. The human language technology (HLT) society in Europe has been particularly zealous for the standardization, making a series of attempts such as EAGLES[1], PAROLE/SIMPLE (Lenci et al., 2000), ISLE/MILE (Calzolari et al., 2003) and LIRICS[2]. These continuous efforts has been crystallized as activities in ISO-TC37/SC4 which aims to make an international standard for language resources.

[1] http://www.ilc.cnr.it/Eagles96/home.html

[2] lirics.loria.fr/documents.html

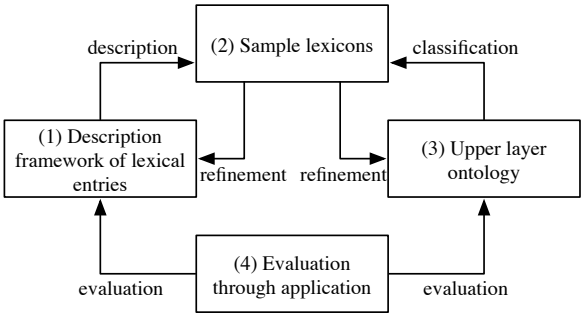


Figure 1: Relations among research items

On the other hand, since Asia has great linguistic and cultural diversity, Asian language resources have received much less attention than their western counterparts. Creating a common standard for Asian language resources that is compatible with an international standard has at least three strong advantages: to increase the competitive edge of Asian countries, to bring Asian countries to closer to their western counterparts, and to bring more cohesion among Asian countries.

To achieve this goal, we have launched a two year project to create a common standard for Asian language resources. The project is comprised of the following four research items.

(1) building a description framework of lexical entries

(2) building sample lexicons

(3) building an upper-layer ontology

(4) evaluating the proposed framework through an application

Figure 1 illustrates the relations among these research items.

Our main aim is the research item (1), building a description framework of lexical entries which

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 827–834,
Sydney, July 2006.

fits with as many Asian languages as possible, and contributing to the ISO-TC37/SC4 activities. As a starting point, we employ an existing description framework, the MILE framework (Bertagna et al., 2004a), to describe several lexical entries of several Asian languages. Through building sample lexicons (research item (2)), we will find problems of the existing framework, and extend it so as to fit with Asian languages. In this extension, we need to be careful in keeping consistency with the existing framework. We start with Chinese, Japanese and Thai as target Asian languages and plan to expand the coverage of languages. The research items (2) and (3) also comprise the similar feedback loop. Through building sample lexicons, we refine an upper-layer ontology. An application built in the research item (4) is dedicated to evaluating the proposed framework. We plan to build an information retrieval system using a lexicon built by extending the sample lexicon.

In what follows, section 2 briefly reviews the MILE framework which is a basis of our description framework. Since the MILE framework is originally designed for European languages, it does not always fit with Asian languages. We exemplify some of the problems in section 3 and suggest some directions to solve them. We expect that further problems will come into clear view through building sample lexicons. Section 4 describes a criteria to choose lexical entries in sample lexicons. Section 5 describes an approach to build an upper-layer ontology which can be sharable among languages. Section 6 describes an application through which we evaluate the proposed framework.

2 The MILE framework for interoperability of lexicons

The ISLE (International Standards for Language Engineering) Computational Lexicon Working Group has consensually defined the MILE (Multilingual ISLE Lexical Entry) as a standardized infrastructure to develop multilingual lexical resources for HLT applications, with particular attention to Machine Translation (MT) and Crosslingual Information Retrieval (CLIR) application systems.

The MILE is a general architecture devised for the encoding of multilingual lexical information, a meta-entry acting as a common representational layer for multilingual lexicons, by allowing integration and interoperability between different monolingual lexicons[3].

This formal and standardized framework to encode MILE-conformant lexical entries is provided to lexicon and application developers by the overall MILE Lexical Model (MLM). As concerns the horizontal organization, the MLM consists of two independent, but interlinked primary components, the monolingual and the multilingual modules. The monolingual component, on the vertical dimension, is organized over three different representational layers which allow to describe different dimensions of lexical entries, namely the morphological, syntactic and semantic layers. Moreover, an intermediate module allows to define mechanisms of linkage and mapping between the syntactic and semantic layers. Within each layer, a basic linguistic information unit is identified; basic units are separated but still interlinked each other across the different layers.

Within each of the MLM layers, different types of lexical object are distinguished :

- the MILE Lexical Classes (MLC) represent the main building blocks which formalize the basic lexical notions. They can be seen as a set of structural elements organized in a layered fashion: they constitute an ontology of lexical objects as an abstraction over different lexical models and architectures. These elements are the backbone of the structural model. In the MLM a definition of the classes is provided together with their attributes and the way they relate to each other. Classes represent notions like InflectionalParadigm, SyntacticFunction, SyntacticPhrase, Predicate, Argument,

- the MILE Data Categories (MDC) which constitute the attributes and values to adorn the structural classes and allow concrete entries to be instantiated. MDC can belong to a shared repository or be user-defined. “NP” and “VP” are data category instances of the class SyntacticPhrase, whereas and “subj” and “obj” are data category instances of the class SyntacticFunction.

- lexical operations, which are special lexical entities allowing the user to define multilin-

[3]MILE is based on the experience derived from existing computational lexicons (e.g. LE-PAROLE, SIMPLE, EuroWordNet, etc.).

gual conditions and perform operations on lexical entries.

Originally, in order to meet expectations placed upon lexicons as critical resources for content processing in the Semantic Web, the MILE syntactic and semantic lexical objects have been formalized in RDF(S), thus providing a web-based means to implement the MILE architecture and allowing for encoding individual lexical entries as instances of the model (Ide et al., 2003; Bertagna et al., 2004b). In the framework of our project, by situating our work in the context of W3C standards and relying on standardized technologies underlying this community, the original RDF schema for ISLE lexical entries has been made compliant to OWL. The whole data model has been formalized in OWL by using Protégé 3.2 beta and has been extended to cover the morphological component as well (see Figure 2). Protégé 3.2 beta has been also used as a tool to instantiate the lexical entries of our sample monolingual lexicons, thus ensuring adherence to the model, encoding coherence and inter- and intra-lexicon consistency.

3 Existing problems with the MILE framework for Asian languages

In this section, we will explain some problematic phenomena of Asian languages and discuss possible extensions of the MILE framework to solve them.

Inflection The MILE provides the powerful framework to describe the information about inflection. **InflectedForm** class is devoted to describe inflected forms of a word, while **InflectionalParadigm** to define general inflection rules. However, there is no inflection in several Asian languages, such as Chinese and Thai. For these languages, we do not use the Inflected Form and Inflectional Paradigm.

Classifier Many Asian languages, such as Japanese, Chinese, Thai and Korean, do not distinguish singularity and plurality of nouns, but use classifiers to denote the number of objects. The followings are examples of classifiers of Japanese.

- *inu* *ni* *hiki* $\cdots$ two dogs
  (dog) (two) (CL)
- *hon* *go* *satsu* $\cdots$ five books
  (book) (five) (CL)

"CL" stands for a classifier. They always follow cardinal numbers in Japanese. Note that different classifiers are used for different nouns. In the above examples, classifier "*hiki*" is used to count noun "*inu* (dog)", while "*satsu*" for "*hon* (book)". The classifier is determined based on the semantic type of the noun.

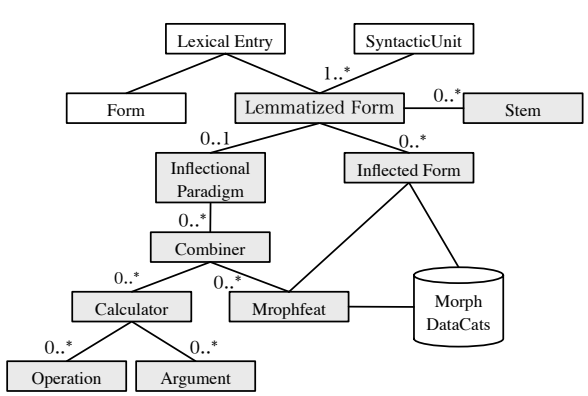
In the Thai language, classifiers are used in various situations (Sornlertlamvanich et al., 1994). The classifier plays an important role in construction with noun to express ordinal, pronoun, for instance. The classifier phrase is syntactically generated according to a specific pattern. Here are some usages of classifiers and their syntactic patterns.

- Enumeration
  (Noun/Verb)-(cardinal number)-(CL)
  e.g. *nakrian* *3* *khon* $\cdots$ three students
  (student) (CL)
- Ordinal
  (Noun)-(CL)-/thi:/-(cardinal number)
  e.g. *kaew* *bai* *thi: 4* $\cdots$ the 4th glass
  (glass) (CL) (4th)
- Determination
  (Noun)-(CL)-(Determiner)
  e.g. *kruangkhidlek* *kruang* *nii*
  (calculator) (CL) (this)
  $\cdots$ this calculator

Classifiers could be dealt as a class of the part-of-speech. However, since classifiers depend on the semantic type of nouns, we need to refer to semantic features in the morphological layer, and vice versa. Some mechanism to link between features beyond layers needs to be introduced into the current MILE framework.

Orthographic variants Many Chinese words have orthographic variants. For instance, the concept of rising can be represented by either character variants of sheng1: 升 or 昇. However, the free variants become non-free in certain compound forms. For instance, only 升 allowed for 公升 'liter', and only 昇 is allowed for 昇華 'to sublime'. The interaction of lemmas and orthographic variations is not yet represented in MILE.

Reduplication as a derivational process In some Asian languages, reduplication of words derives another word, and the derived word often has a different part-of-speech. Here are some examples of reduplication in Chinese. Man4 慢 'to be slow' is a state verb, while a reduplicated form




```
<LemmatizedForm rdf:ID="LFstar">
 <hasInflectedForm>
  <InflectedForm rdf:ID="stars">
   <hasMorphoFeat>
  <MorphoFeat rdf:ID="pl">
   <number rdf:datatype="http://www.w3c.org/
2001/ XMLSchema#string">
     plural
   </number>
  </MorphoFeat>
   </hasMorphoFeat>
  </InflectedForm>
 </hasInflectedForm>
 <hasInflectedForm>
  <InflectedForm rdf:ID="star">
   <hasMorphoFeat>
  <MorphoFeat rdf:ID="sg">
   <number rdf:datatype="http://www.w3c.org/
2001/ XMLSchema#string">
     singular
   </number>
  </MorphoFeat>
   </hasMorphoFeat>
  </InflectedForm>
 </hasInflectedForm>
</LemmatiedForm>
```

Figure 2: Formalization of the morphological layer and excerpt of a sample RDF instantiation

man4-man4 慢慢 is an adverb. Another example of reduplication involves verbal aspect. Kan4 看 'to look' is an activity verb, while the reduplicative form kan4-kan4 看看, refers to the tentative aspect, introducing either stage-like sub-division or the event or tentativeness of the action of the agent. This morphological process is not provided for in the current MILE standard.

There are also various usages of reduplication in Thai. Some words reduplicate themselves to add a specific aspect to the original meaning. The reduplication can be grouped into 3 types according to the tonal sound change of the original word.

- Word reduplication without sound change
  e.g. /dek-dek/ · · · (N) children, (ADV) childishly, (ADJ) childish
  /sa:w-sa:w/ · · · (N) women
- Word reduplication with high tone on the first word
  e.g. /dam4-dam/ · · · (ADJ) extremely black
  /bo:i4-bo:i/ · · · (ADV) really often
- Triple word reduplication with high tone on the second word
  e.g. /dern-dern4-dern/ ·· (V) intensively walk
  /norn-norn4-norn/··(V) intensively sleep

In fact, only the reduplication of the same sound is accepted in the written text, and a special symbol, namely /mai-yamok/ is attached to the original word to represent the reduplication. The reduplication occurs in many parts-of-speech, such as noun, verb, adverb, classifier, adjective, preposition. Furthermore, various aspects can be added to the original meaning of the word by reduplication, such as pluralization, emphasis, generalization, and so on. These aspects should be instantiated as features.

Change of parts-of-speech by affixes Affixes change parts-of-speech of words in Thai (Charoenporn et al., 1997). There are three prefixes changing the part-of-speech of the original word, namely /ka:n/, /khwa:m/, /ya:ng/. They are used in the following cases.

- Nominalization
  /ka:n/ is used to prefix an action verb and /khwa:m/ is used to prefix a state verb in nominalization such as /ka:n-tham-nga:n/ (working), /khwa:m-suk/ (happiness).
- Adverbialization
  An adverb can be derived by using /ya:ng/ to prefix a state verb such as /ya:ng-di:/ (well).

Note that these prefixes are also words, and form multi-word expressions with the original word. This phenomenon is similar to derivation which is not handled in the current MILE framework. Derivation is traditionally considered as a different phenomenon from inflection, and current MILE focuses on inflection. The MILE framework is already being extended to treat such linguistic phenomenon, since it is important to European languages as well. It would be handled in either the morphological layer or syntactic layer.

Function Type Function types of predicates (verbs, adjectives etc.) might be handled in a partially different way for Japanese. In the syntactic layer of the MILE framework, **FunctionType** class is prepared to denote subcategorization frames of predicates, and they have function types such as "subj" and "obj". For example, the verb "eat" has two **FunctionType** data categories of "subj" and "obj". Function types basically stand for positions of case filler nouns. In Japanese, cases are usually marked by postpositions and case filler positions themselves do not provide much information on case marking. For example, both of the following sentences mean the same, "She eats a pizza."

- *kanojo ga piza wo taberu*
  (she) (NOM) (pizza) (ACC) (eat)
- *piza wo kanojo ga taberu*
  (pizza) (ACC) (she) (NOM) (eat)

"*Ga*" and "*wo*" are postpositions which mark nominative and accusative cases respectively. Note that two case filler nouns "she" and "pizza" can be exchanged. That is, the number of slots is important, but their order is not.

For Japanese, we might use the set of postpositions as values of **FunctionType** instead of conventional function types such as "subj" and "obj". It might be an user defined data category or language dependent data category. Furthermore, it is preferable to prepare the mapping between Japanese postpositions and conventional function types. This is interesting because it seems more a terminological difference, but the model can be applied also to Japanese.

4 Building sample lexicons

4.1 Swadesh list and basic lexicon

The issue involved in defining a basic lexicon for a given language is more complicated than one may think (Zhang et al., 2004). The naive approach of simply taking the most frequent words in a language is flawed in many ways. First, all frequency counts are corpus-based and hence inherit the bias of corpus sampling. For instance, since it is easier to sample written formal texts, words used predominantly in informal contexts are usually underrepresented. Second, frequency of content words is topic-dependent and may vary from corpus to corpus. Last, and most crucially, frequency of a word does not correlate to its conceptual necessity, which should be an important, if not only, criteria for core lexicon. The definition of a cross-lingual basic lexicon is even more complicated. The first issue involves determination of cross-lingual lexical equivalencies. That is, how to determine that word *a* (and not *a'*) in language *A* really is word *b* in language *B*. The second issue involves the determination of what is a basic word in a multilingual context. In this case, not even the frequency offers an easy answer since lexical frequency may vary greatly among different languages. The third issue involves lexical gaps. That is, if there is a word that meets all criteria of being a basic word in language *A*, yet it does not exist in language *D* (though it may exist in languages *B*, and *C*). Is this word still qualified to be included in the multilingual basic lexicon?

It is clear not all the above issues can be unequivocally solved with the time frame of our project. Fortunately, there is an empirical core lexicon that we can adopt as a starting point. The Swadesh list was proposed by the historical linguist Morris Swadesh (Swadesh, 1952), and has been widely used by field and historical linguists for languages over the world. The Swadesh list was first proposed as lexico-statistical metrics. That is, these are words that can be reliably expected to occur in all historical languages and can be used as the metrics for quantifying language variations and language distance. The Swadesh list is also widely used by field linguists when they encounter a new language, since almost all of these terms can be expected to occur in any language. Note that the Swadesh list consists of terms that embody human direct experience, with culture-specific terms avoided. Swadesh started with a 215 items list, before cutting back to 200 items and then to 100 items. A standard list of 207 items is arrived at by unifying the 200 items list and the 100 items list. We take the 207 terms from the Swadesh list as the core of our basic lexicon. Inclusion of the Swadesh list also gives us the possibility of covering many Asian languages in which we do not have the resources to make a full and fully annotated lexicon. For some of these languages, a Swadesh lexicon for reference is provided by a collaborator.

4.2 Aligning multilingual lexical entries

Since our goal is to build a multilingual sample lexicon, it is required to align words in several

Asian languages. In this subsection, we propose a simple method to align words in different languages. The basic idea for multilingual alignment is an intermediary by English. That is, first we prepare word pairs between English and other languages, then combine them together to make correspondence among words in several languages. The multilingual alignment method currently we consider is as follows:

1. Preparing the set of frequent words of each language

   Suppose that $\{Jw_i\}$, $\{Cw_i\}$, $\{Tw_i\}$ is the set of frequent words of Japanese, Chinese and Thai, respectively. Now we try to construct a multilingual lexicon for these three languages, however, our multilingual alignment method can be easily extended to handle more languages.

2. Obtaining English translations

   A word Xw_i is translated into a set of English words EXw_{ij} by referring to the bilingual dictionary, where X denotes one of our languages, J, C or T. We can obtain mappings as in (1).

$$\begin{array}{ll} Jw_1: & EJw_{11}, EJw_{12}, \cdots \\ Jw_2: & EJw_{21}, EJw_{22}, \cdots \\ \vdots & \\ \hline Cw_1: & ECw_{11}, ECw_{12}, \cdots \\ Cw_2: & ECw_{21}, ECw_{22}, \cdots \\ \vdots & \\ \hline Tw_1: & ETw_{11}, ETw_{12}, \cdots \\ Tw_2: & ETw_{21}, ETw_{22}, \cdots \\ \vdots & \end{array} \quad (1)$$

   Notice that this procedure is automatically done and ambiguities would be left at this stage.

3. Generating new mapping

   From mappings in (1), a new mapping is generated by inverting the key. That is, in the new mapping, a key is an English word Ew_i and a correspondence for each key is sets of translations XEw_{ij} for 3 languages, as shown in (2):

$$\begin{array}{ll} Ew_1: & (JEw_{11}, JEw_{12}, \cdots) \\ & (CEw_{11}, CEw_{12}, \cdots) \\ & (TEw_{11}, TEw_{12}, \cdots) \\ Ew_2: & (JEw_{21}, JEw_{22}, \cdots) \\ & (CEw_{21}, CEw_{22}, \cdots) \\ & (TEw_{21}, TEw_{22}, \cdots) \\ \vdots & \end{array} \quad (2)$$

   Notice that at this stage, correspondence between different languages is very loose, since they are aligned on the basis of sharing only a single English word.

4. Refinement of alignment

   Groups of English words are constructed by referring to the WordNet synset information. For example, suppose that Ew_i and Ew_j belong to the same synset S_k. We will make a new alignment by making an intersection of $\{XEw_i\}$ and $\{XEw_j\}$ as shown in (3).

$$\begin{array}{r} Ew_i: (JEw_{i1}, \cdot\cdot)\,(CEw_{i1}, \cdot\cdot)\,(TEw_{i1}, \cdot\cdot) \\ Ew_j: (JEw_{j1}, \cdot\cdot)\,(CEw_{j1}, \cdot\cdot)\,(TEw_{j1}, \cdot\cdot) \\ \Downarrow \text{intersection} \\ S_k: \ (JEw'_{k1}, \cdot\cdot)(CEw'_{k1}, \cdot\cdot)(TEw'_{k1}, \cdot\cdot) \end{array} \quad (3)$$

   In (3), the key is a synset S_k, which is supposed to be a conjunction of Ew_i and Ew_j, and the counterpart is the intersection of set of translations for each language. This operation would reduce the number of words of each language. That means, we can expect that the correspondence among words of different languages becomes more precise. This new word alignment based on a synset is a final result.

To evaluate the performance of this method, we conducted a preliminary experiment using the Swadesh list. Given the Swadesh list of Chinese, Italian, Japanese and Thai as a gold standard, we tried to replicate these lists from the English Swadesh list and bilingual dictionaries between English and these languages. In this experiment, we did not perform the refinement step with WordNet. From 207 words in the Swadesh list, we dropped 4 words ("at", "in", "with" and "and") due to their too many ambiguities in translation.

As a result, we obtained 181 word groups aligned across 5 languages (Chinese, English, Italian, Japanese and Thai) for 203 words. An aligned word group was judged "correct" when the words of each language include only words in the Swadesh list of that language. It was judged "partially correct" when the words of a language also include the words which are not in the Swadesh list. Based on the correct instances, we obtain 0.497 for precision and 0.443 for recall. These figures go up to 0.912 for precision and 0.813 for recall when based on the partially correct instances. This is quite a promising result.

5 Upper-layer ontology

The empirical success of the Swadesh list poses an interesting question that has not been explored before. That is, does the Swadesh list instantiates a shared, fundamental human conceptual structure? And if there is such as a structure, can we discover it?

In the project these fundamental issues are associated with our quest for cross-lingual interoperability. We must make sure that the items of the basic lexicon are given the same interpretation. One measure taken to ensure this consists in constructing an upper-ontology based on the basic lexicon. Our preliminary work of mapping the Swadesh list items to SUMO (Suggested Upper Merged Ontology) (Niles and Pease, 2001) has already been completed. We are in the process of mapping the list to DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Masolo et al., 2003). After the initial mapping, we carry on the work to restructure the mapped nodes to form a genuine conceptual ontology based on the language universal basic lexical items. However one important observation that we have made so far is that the success of the Swadesh list is partly due to its underspecification and to the liberty it gives to compilers of the list in a new language. If this idea of underspecification is essential for basic lexicon for human languages, then we must resolve this apparent dilemma of specifying them in a formal ontology that requires fully specified categories. For the time being, genuine ambiguities resulted in the introduction of each disambiguated sense in the ontology. We are currently investigating another solution that allows the inclusion of underspecified elements in the ontology without threatening its coherence. More specifically we introduce a underspecified relation in the structure for linking the underspecified meaning to the different specified meaning. The specified meanings are included in the taxonomic hierarchy in a traditional manner, while a hierarchy of underspecified meanings can be derived thanks to the new relation. An underspecified node only inherits from the most specific common mother of its fully specified terms. Such distinction avoids the classical misuse of the subsumption relation for representing multiple meanings. This method does not reflect a dubious collapse of the linguistic and conceptual levels but the treatment of such underspecifications as truly conceptual. Moreover we hope this proposal will provide a knowledge representation framework for the multilingual alignment method presented in the previous section.

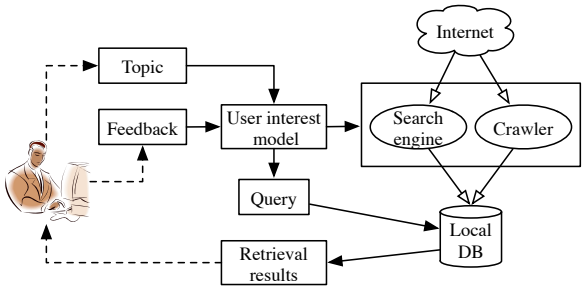


Figure 3: The system architecture

Finally, our ontology will not only play the role of a structured interlingual index. It will also serve as a common conceptual base for lexical expansion, as well as for comparative studies of the lexical differences of different languages.

6 Evaluation through an application

To evaluate the proposed framework, we are building an information retrieval system. Figure 3 shows the system architecture.

A user can input a topic to retrieve the documents related to that topic. A topic can consist of keywords, website URL's and documents which describe the topic. From the topic information, the system builds a user interest model. The system then uses a search engine and a crawler to search for information related to this topic in WWW and stores the results in the local database. Generally, the search results include many noises. To filter out these noises, we build a query from the user interest model and then use this query to retrieve documents in the local database. Those documents similar to the query are considered as more related to the topic and the user's interest, and are returned to the user. When the user obtains these retrieval results, he can evaluate these documents and give the feedback to the system, which is used for the further refinement of the user interest model.

Language resources can contribute to improving the system performance in various ways. Query expansion is a well-known technique which expands user's query terms into a set of similar and related terms by referring to ontologies. Our system is based on the vector space model (VSM) and traditional query expansion can be applicable using the ontology.

There has been less research on using lexical in-

formation for information retrieval systems. One possibility we are considering is query expansion by using predicate-argument structures of terms. Suppose a user inputs two keywords, "hockey" and "ticket" as a query. The conventional query expansion technique expands these keywords to a set of similar words based on an ontology. By referring to predicate-argument structures in the lexicon, we can derive actions and events as well which take these words as arguments. In the above example, by referring to the predicate-argument structure of "buy" or "sell", and knowing that these verbs can take "ticket" in their object role, we can add "buy" and "sell" to the user's query. This new type of expansion requires rich lexical information such as predicate argument structures, and the information retrieval system would be a good touchstone of the lexical information.

7 Concluding remarks

This paper outlined a new project for creating a common standard for Asian language resources in cooperation with other initiatives. We start with three Asian languages, Chinese, Japanese and Thai, on top of the existing framework which was designed mainly for European languages. We plan to distribute our draft to HLT societies of other Asian languages, requesting for their feedback through various networks, such as the Asian language resource committee network under Asian Federation of Natural Language Processing (AFNLP)[4], and Asian Language Resource Network project[5]. We believe our efforts contribute to international activities like ISO-TC37/SC4[6] (Francopoulo et al., 2006) and to the revision of the ISO Data Category Registry (ISO 12620), making it possible to come close to the ideal international standard of language resources.

Acknowledgment

This research was carried out through financial support provided under the NEDO International Joint Research Grant Program (NEDO Grant).

References

F. Bertagna, A. Lenci, M. Monachini, and N. Calzolari. 2004a. Content interoperability of lexical resources, open issues and "MILE" perspectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, pages 131–134.

F. Bertagna, A. Lenci, M. Monachini, and N. Calzolari. 2004b. The MILE lexical classes: Data categories for content interoperability among lexicons. In *A Registry of Linguistic Data Categories within an Integrated Language Resources Repository Area – LREC2004 Satellite Workshop*, page 8.

N. Calzolari, F. Bertagna, A. Lenci, and M. Monachini. 2003. Standards and best practice for multilingual computational lexicons. MILE (the multilingual ISLE lexical entry). ISLE Deliverable D2.2&3.2.

T. Charoenporn, V. Sornlertlamvanich, and H. Isahara. 1997. Building a large Thai text corpus — part-of-speech tagged corpus: ORCHID—. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*.

G. Francopoulo, G. Monte, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. 2006. Lexical markup framework (LMF). In *Proceedings of LREC2006 (forthcoming)*.

N. Ide, A. Lenci, and N. Calzolari. 2003. RDF instantiation of ISLE/MILE lexical entries. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 25–34.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowsky, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography, Special Issue, Dictionaries, Thesauri and Lexical-Semantic Relations*, XIII(4):249–263.

C. Masolo, A. Borgo, S.; Gangemi, N. Guarino, and A. Oltramari. 2003. Wonderweb deliverable d18 –ontology library (final)–. Technical report, Laboratory for Applied Ontology, ISTC-CNR.

I. Niles and A Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

V. Sornlertlamvanich, W. Pantachat, and S. Meknavin. 1994. Classifier assignment by corpus-based approach. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 556–561.

M. Swadesh. 1952. Lexico-statistical dating of prehistoric ethnic contacts: With special reference to north American Indians and Eskimos. In *Proceedings of the American Philo-sophical Society*, volume 96, pages 452–463.

H. Zhang, C. Huang, and S. Yu. 2004. Distributional consistency: A general method for defining a core lexicon. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, pages 1119–1222.

[4] http://www.afnlp.org/

[5] http://www.language-resource.net/

[6] http://www.tc37sc4.org/