Query Expansion using LMF-Compliant Lexical Resources

Tokunaga Takenobu, *Tokyo Inst. of Tech.*
Dain Kaplan, *Tokyo Inst. of Tech.*
Nicoletta Calzolari, *ILC/CNR*
Monica Monachini, *ILC/CNR*
Claudia Soria, *ILC/CNR*
Virach Sornlertlamvanich, *TCL, NICT*
Thatsanee Charoenporn, *TCL, NICT*
Xia Yingju, *Fujitsu R&D Center*
Chu-Ren Huang, *The Hong Kong Polytec. Univ.*
Shu-Kai Hsieh, *National Taiwan Normal Univ.*
Shirai Kiyoaki, *JAIST*

Abstract

This paper reports prototype multilingual query expansion system relying on LMF compliant lexical resources. The system is one of the deliverables of a three-year project aiming at establishing an international standard for language resources which is applicable to Asian languages. Our important contributions to ISO 24613, standard Lexical Markup Framework (LMF) include its robustness to deal with Asian languages, and its applicability to cross-lingual query tasks, as illustrated by the prototype introduced in this paper.

1 Introduction

During the last two decades corpus-based approaches have come to the forefront of NLP research. Since without corpora there can be no corpus-based research, the creation of such language resources has also necessarily advanced as well, in a mutually beneficial synergetic relationship. One of the advantages of corpus-based approaches is that the techniques used are less language specific than classical rule-based approaches where a human analyses the behaviour of target languages and constructs rules manually. This naturally led the way for international resource standardisation, and indeed there is a long standing precedent in the West for it. The Human Language Technology (HLT) society in Europe has been particularly zealous in this regard, propelling the creation of resource interoperability through a series of initiatives, namely EAGLES (Sanfilippo et al., 1999), PAROLE/SIMPLE (Lenci et al., 2000), ISLE/MILE (Ide et al., 2003), and LIRICS[1]. These continuous efforts have matured into activities in ISO-TC37/SC4[2], which aims at making an international standard for language resources.

However, due to the great diversity of languages themselves and the differing degree of technological development for each, Asian languages, have received less attention for creating resources than their Western counterparts. Thus, it has yet to be determined if corpus-based techniques developed for well-computerised languages are applicable on a broader scale to all languages. In order to efficiently develop Asian language resources, utilising an international standard in this creation has substantial merits.

We launched a three-year project to create an international standard for language resources that includes Asian languages. We took the following approach in seeking this goal.

- Based on existing description frameworks, each research member tries to describe several lexical entries and find problems with them.
- Through periodical meetings, we exchange information about problems found and generalise them to propose solutions.
- Through an implementation of an application system, we verify the effectiveness of the proposed framework.

Below we summarise our significant contribution to an International Standard (ISO24613; Lexical Markup Framework: LMF).

1st year After considering many characteristics of Asian languages, we elucidated the shortcomings of the LMF draft (ISO24613 Rev.9). The draft lacks the following devices for Asian languages.

[1] http://lirics.loria.fr/

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

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 145–152,
Suntec, Singapore, 6-7 August 2009.

(1) A mapping mechanism between syntactic and semantic arguments
(2) Derivation (including reduplication)
(3) Classifiers
(4) Orthography
(5) Honorifics

Among these, we proposed solutions for (1) and (2) to the ISO-TC37 SC4 working group.

2nd year We proposed solutions for above the (2), (3) and (4) in the comments of the Committee Draft (ISO24613 Rev. 13) to the ISO-TC37 SC4 working group. Our proposal was included in DIS (Draft International Standard).

(2') a package for derivational morphology
(3') the syntax-semantic interface resolving the problem of classifiers
(4') representational issues with the richness of writing systems in Asian languages

3rd year Since ISO 24613 was in the FDIS stage and fairly stable, we built sample lexicons in Chinese, English, Italian, Japanese, and Thai based on ISO24613. At the same time, we implemented a query expansion system utilising rich linguistic resources including lexicons described in the ISO 24613 framework. We confirmed that a system was feasible which worked on the tested languages (including both Western and Asian languages) when given lexicons compliant with the framework. ISO 24613 (LMF) was approved by the October 2008 ballot and published as ISO-24613:2008 on 17th November 2008.

Since we have already reported our first 2 year activities elsewhere (Tokunaga and others, 2006; Tokunaga and others, 2008), we focus on the above query expansion system in this paper.

2 Query expansion using LMF-compliant lexical resources

We evaluated the effectiveness of LMF on a multilingual information retrieval system, particularly the effectiveness for linguistically motivated query expansion.

The linguistically motivated query expansion system aims to refine a user's query by exploiting the richer information contained within a lexicon described using the adapted LMF framework. Our lexicons are completely complaint with this international standard. For example, a user inputs a keyword "ticket" as a query. Conventional query expansion techniques expand this keyword to a set of related words by using thesauri or ontologies (Baeza-Yates and Ribeiro-Neto, 1999). Using the framework proposed by this project, expanding the user's query becomes a matter of following links within the lexicon, from the source lexical entry or entries through predicate-argument structures to all relevant entries (Figure 1). We focus on expanding the user inputted list of nouns to relevant verbs, but the reverse would also be possible using the same technique and the same lexicon. This link between entries is established through the *semantic type* of a given sense within a lexical entry. These semantic types are defined by higher-level ontologies, such as MILO or SIMPLE (Lenci et al., 2000) and are used in semantic predicates that take such semantic types as a restriction argument. Since senses for verbs contain a link to a semantic predicate, using this semantic type, the system can then find any/all entries within the lexicon that have this semantic type as the value of the restriction feature of a semantic predicate for any of their senses. As a concrete example, let us continue using the "ticket" scenario from above. The lexical entry for "ticket" might contain a semantic type definition something like in Figure 2.

```
<LexicalEntry ...>
  <feat att="POS" val="N"/>
  <Lemma>
    <feat att="writtenForm"
          val="ticket"/>
  </Lemma>
  <Sense ...>
    <feat att="semanticType"
          val="ARTIFACT"/>
    ...
  </Sense>
  ...
</LexicalEntry>
```

Figure 2: Lexical entry for "ticket"

By referring to the lexicon, we can then derive any actions and events that take the semantic type "ARTIFACT" as an argument.

First all semantic predicates are searched for arguments that have an appropriate restriction, in this case "ARTIFACT" as shown in Figure 3, and then any lexical entries that refer to these predicates are returned. An equally similar definition would exist for "buy", "find" and so on. Thus, by referring to the predicate-argument structure of related verbs, we know that these verbs can take

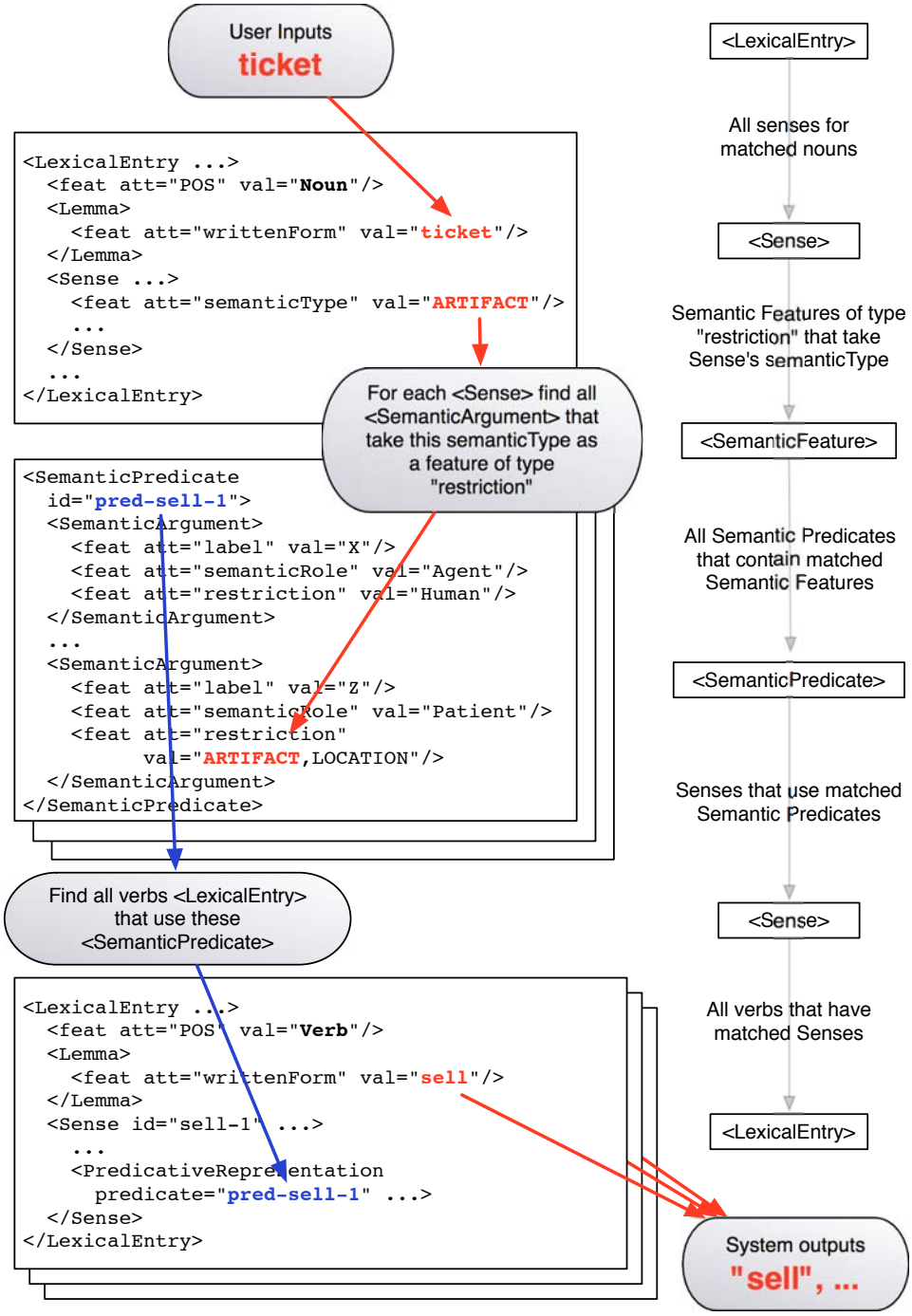

Figure 1: QE Process Flow

```
<LexicalEntry ...>
  <feat att="POS" val="V"/>
  <Lemma>
    <feat att="writtenForm"
          val="sell"/>
  </Lemma>
  <Sense id="sell-1" ...>
    <feat att="semanticType"
          val="Transaction"/>
    <PredicativeRepresentation
      predicate="pred-sell-1"
      correspondences="map-sell1">
  </Sense>
</LexicalEntry>

<SemanticPredicate id="pred-sell-1">
  <SemanticArgument ...>
    ...
    <feat att="restriction"
          val="ARTIFACT"/>
  </SemanticArgument>
</SemanticPredicate>
```

Figure 3: Lexical entry for "sell" with its semantic predicate

"ticket" in the role of object. The system then returns all relevant entries, here "buy", "sell" and "find", in response to the user's query. Figure 1 schematically shows this flow.

3 A prototype system in detail

3.1 Overview

To test the efficacy of the LMF-compliant lexical resources, we created a system implementing the query expansion mechanism explained above. The system was developed in Java for its "compile once, run anywhere" portability and its high-availability of reusable off-the-shelf components. On top of Java 5, the system was developed using JBoss Application Server 4.2.3, the latest standard, stable version of the product at the time of development. To provide fast access times, and easy traversal of relational data, a RDB was used. The most popular free open-source database was selected, MySQL, to store all lexicons imported into the system, and the system was accessed, as a web-application, via any web browser.

3.2 Database

The finalised database schema is shown in Figure 4. It describes the relationships between entities, and more or less mirrors the classes found within the adapted LMF framework, with mostly only minor exceptions where it was efficacious for querying the data. Due to space constraints, metadata fields, such as creation time-stamps have been left out of this diagram. Since the system also allows for multiple lexicons to co-exist, a *lexicon_id* resides in every table. This foreign key has been highlighted in a different color, but not connected via arrows to make the diagram easier to read. In addition, though in actuality this foreign key is not required for all tables, it has been inserted as a convenience for querying data more efficiently, even within join tables (indicated in blue). Having multiple lexical resources co-existing within the same database allows for several advantageous features, and will be described later. Some tables also contain a *text_id*, which stores the original id attribute for that element found within the XML. This is not used in the system itself, and is stored only for reference.

3.3 System design

As mentioned above, the application is deployed to JBoss AS as an *ear*-file. The system itself is composed of java classes encapsulating the data contained within the database, a Parsing/Importing class for handling the LMF XML files after they have been validated, and JSPs, which contain HTML, for displaying the interface to the user. There are three main sections to the application: Search, Browse, and Configure. Explaining last to first, the Configure section, shown in Figure 5, allows users to create a new lexicon within the system or append to an existing lexicon by uploading a LMF XML file from their web browser, or delete existing lexicons that are no longer needed/used. After import, the data may be immediately queried upon with no other changes to system configuration, from within both the Browse and Search sections. Regardless of language, the rich syntactic/semantic information contained within the lexicon is sufficient for carrying out query expansion on its own.

The Browse section (Figure 6) allows the user to select any available lexicon to see the relationships contained within it, which contains tabs for viewing all noun to verb connections, a list of nouns, a list of verbs, and a list of semantic types. Each has appropriate links allowing the user to easily jump to a different tab of the system. Clicking on a noun takes them to the Search section (Figure 7). In this section, the user may select many lexicons to perform query extraction on, as is visible in Figure 7.

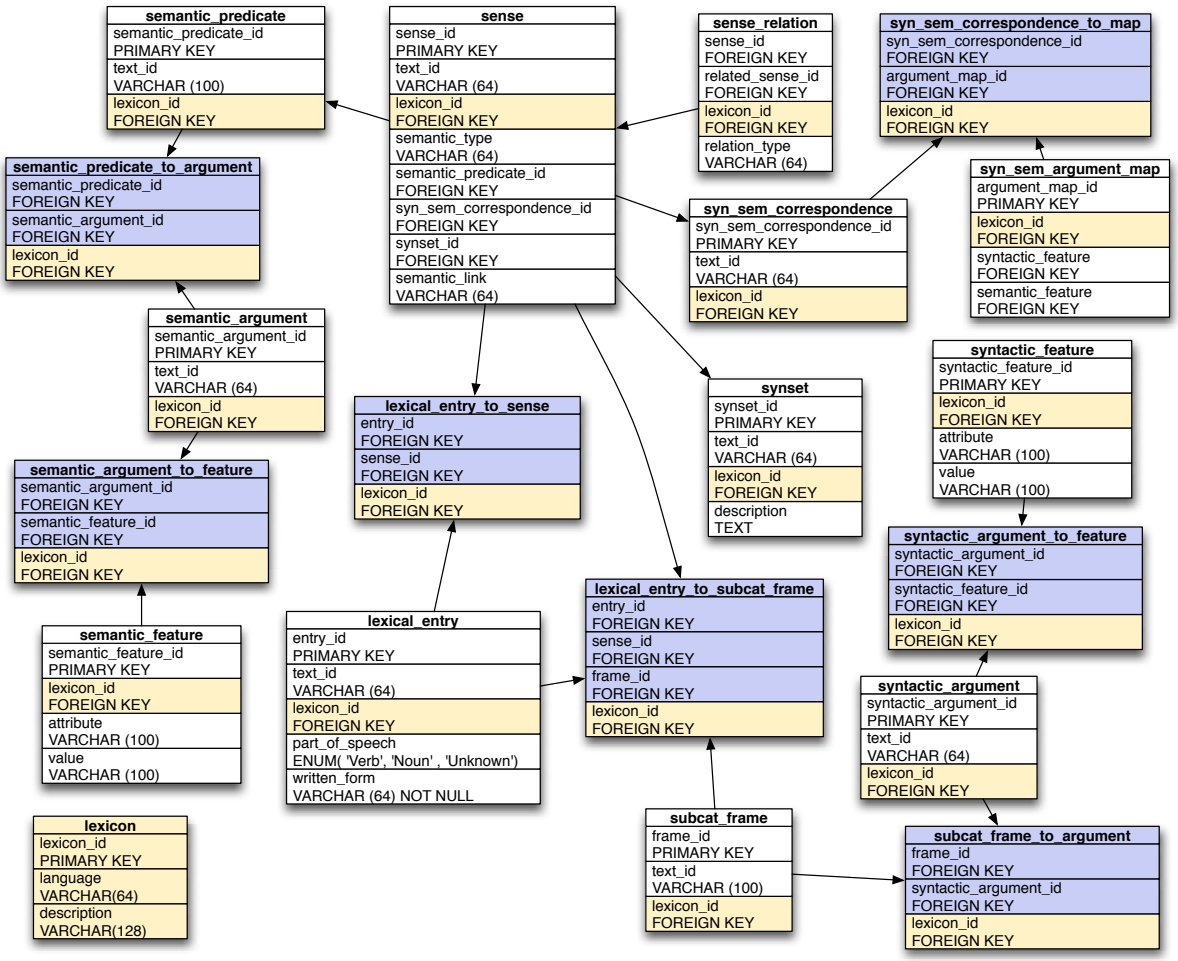

Figure 4: Database schema

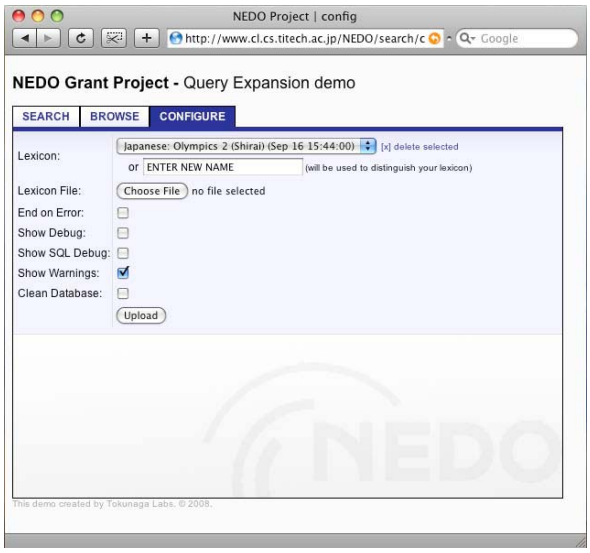

Figure 5: QE System - Configure

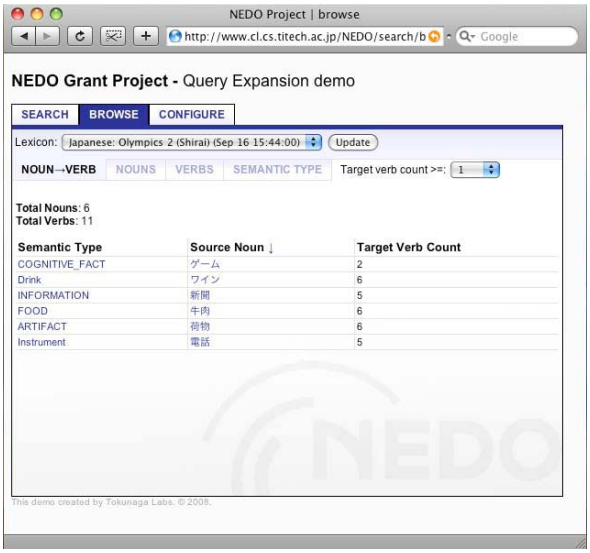

Figure 6: QE System - Browse

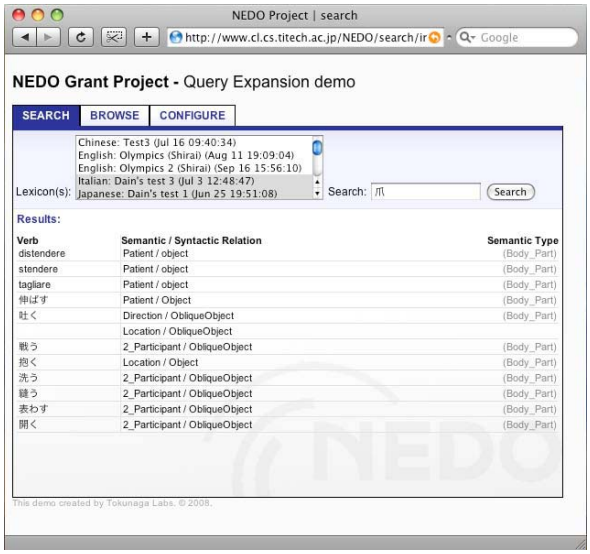

Figure 7: QE System - Search

3.4 Semantic information

This new type of query expansion requires rich lexical information. We augmented our data using the SIMPLE ontology for semantic types, using the same data for different languages. This had the added benefit of allowing *cross*-language expansion as a result. In steps two and three of Figure 1 when senses are retrieved that take specific semantic types as arguments, this process can be done across all (or as many as are selected) lexicons in the database. Thus, results such as are shown in Figure 7 are possible. In this figure the Japanese word for "nail" is entered, and results for both selected languages, Japanese *and* Italian, are returned. This feature requires the unification of the semantic type ontology strata.

3.5 Possible extension

Next steps for the QE platform are to explore the use of other information already defined within the adapted framework, specifically sense relations. Given to the small size of our sample lexicon, data sparsity is naturally an issue, but hopefully by exploring and exploiting these sense relations properly, the system may be able to further expand a user's query to include a broader range of selections using any additional semantic types belonging to these related senses. The framework also contains information about the order in which syntactic arguments should be placed. This information should be used to format the results from the user's query appropriately.

4 An Additional Evaluation

We conducted some additional query expansion experiments using a corpus that was acquired from Chinese LDC (No. "2004-863-009") as a base (see below). This corpus marked an initial achievement in building a multi-lingual parallel corpus for supporting development of cross-lingual NLP applications catering to the Beijing 2008 Olympics.

The corpus contains parallel texts in Chinese, English and Japanese and covers 5 domains that are closely related to the Olympics: traveling, dining, sports, traffic and business. The corpus consists of example sentences, typical dialogues and articles from the Internet, as well as other language teaching materials. To deal with the different languages in a uniform manner, we converted the corpus into our proposed LMF-compliant lexical resources framework, which allowed the system to expand the query between all the languages within the converted resources without additional modifications.

As an example of how this IR system functioned, suppose that Mr. Smith will be visiting Beijing to see the Olympic games and wants to know how to buy a newspaper. Using this system, he would first enter the query "newspaper". For this query, with the given corpus, the system returns 31 documents, fragments of the first 5 shown below.

(1) I'll bring an English *newspaper* immediately.

(2) Would you please hand me the *newspaper*.

(3) There's no use to go over the *newspaper* ads.

(4) Let's consult the *newspaper* for such a film.

(5) I have little confidence in what the *newspapers* say.

Yet it can be seen that the displayed results are not yet useful enough to know how to buy a newspaper, though useful information may in fact be included within some of the 31 documents. Using the lexical resources, the query expansion module suggests "buy", "send", "get", "read", and "sell" as candidates to add for a revised query.

Mr. Smith wants to buy a newspaper, so he selects "buy" as the expansion term. With this query the system returns 11 documents, fragments of the first 5 listed below.

(6) I'd like some *newspapers*, please.

(7) Oh, we have a barber shop, a laundry, a store, telegram services, a *newspaper* stand, table tennis, video games and so on.

(8) We can put an ad in the *newspaper*.

(9) Have you read about the Olympic Games of Table Tennis in today's *newspaper*, Miss?

(10) *newspaper* says we must be cautious about tidal waves.

This list shows improvement, as information about newspapers and shopping is present, but still appears to lack any documents directly related to *how* to buy a newspaper.

Using co-occurrence indexes, the IR system returns document (11) below, because the noun "newspaper" and the verb "buy" appear in the same sentence.

(11) You can make change at some stores, just buy a *newspaper* or something.

From this example it is apparent that this sort of query expansion is still too naive to apply to real IR systems. It should be noted, however, that our current aim of evaluation was in confirming the advantage of LMF in dealing with multiple languages, for which we conducted a similar run with Chinese and Japanese. Results of these tests showed that in following the LMF framework in describing lexical resources, it was possibile to deal with all three languages without changing the mechanics of the system at all.

5 Discussion

LMF is, admittedly, a "high-level" specification, that is, an abstract model that needs to be further developed, adapted and specified by the lexicon encoder. LMF does not provide any off-the-shelf representation for a lexical resource; instead, it gives the basic structural components of a lexicon, leaving full freedom for modeling the particular features of a lexical resource. One drawback is that LMF provides only a specification manual with a few examples. Specifications are by no means instructions, exactly as XML specifications are by no means instructions on how to represent a particular type of data.

Going from LMF specifications to a true instantiation of an LMF-compliant lexicon is a long way, and comprehensive, illustrative and detailed examples for doing this are needed. Our prototype system provides a good starting example for this direction. LMF is often taken as a prescriptive description, and its examples taken as pre-defined normative examples to be used as coding guidelines. Controlled and careful examples of conversion to LMF-compliant formats are also needed to avoid too subjective an interpretation of the standard.

We believe that LMF will be a major base for various SemanticWeb applications because it provides interoperability across languages and directly contributes to the applications themselves, such as multilingual translation, machine aided translation and terminology access in different languages.

From the viewpoint of LMF, our prototype demonstrates the adaptability of LMF to a representation of real-scale lexicons, thus promoting its adoption to a wider community. This project is one of the first test-beds for LMF (as one of its drawbacks being that it has not been tested on a wide variety of lexicons), particularly relevant since it is related to both Western and Asian language lexicons. This project is a concrete attempt to specify an LMF-compliant XML format, tested for representative and parsing efficiency, and to provide guidelines for the implementation of an LMF-compliant format, thus contributing to the reduction of subjectivity in interpretation of standards.

From our viewpoint, LMF has provided a format for exchange of information across differently conceived lexicons. Thus LMF provides a standardised format for relating them to other lexical models, in a linguistically controlled way. This seems an important and promising achievement in order to move the sector forward.

6 Conclusion

This paper described the results of a three-year project for creating an international standard for language resources in cooperation with other initiatives. In particular, we focused on query expansion using the standard.

Our main contribution can be summarised as follows.

- We have contributed to ISO TC37/SC4 activities, by testing and ensuring the portability and applicability of LMF to the development of a description framework for NLP lexicons for Asian languages. Our contribution includes (1) a package for derivational

morphology, (2) the syntax-semantic interface with the problem of classifiers, and (3) representational issues with the richness of writing systems in Asian languages. As of October 2008, LMF including our contributions has been approved as the international standard ISO 26413.

- We discussed Data Categories necessary for Asian languages, and exemplified several Data Categories including reduplication, classifier, honorifics and orthography. We will continue to harmonise our activity with that of ISO TC37/SC4 TDG2 with respect to Data Categories.

- We designed and implemented an evaluation platform of our description framework. We focused on linguistically motivated query expansion module. The system works with lexicons compliant with LMF and ontologies. Its most significant feature is that the system can deal with any language as far as the those lexicons are described according to LMF. To our knowledge, this is the first working system adopting LMF.

In this project, we mainly worked on 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 results 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)[3], and the Asian Language Resource Network project[4]. We believe our efforts contribute to international activities like ISO-TC37/SC4[5] (Francopoulo et al., 2006).

Acknowledgments

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

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

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

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.

A. Sanfilippo, N. Calzolari, S. Ananiadou, R. Gaizauskas, P. Saint-Dizier, and P. Vossen. 1999. EAGLES recommendations on semantic encoding. EAGLES LE3-4244 Final Report.

T. Tokunaga et al. 2006. Infrastructure for standardization of Asian language resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 827–834.

T. Tokunaga et al. 2008. Adapting international standard for asian language technologies. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

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

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

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