Construction of an Infrastructure for Providing Users with Suitable Language Resources

Hitomi Tohyama† **Shunsuke Kozawa**† **Kiyotaka Uchimoto**‡
Shigeki Matsubara† and **Hitoshi Isahara**‡
†Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{hitomi,kozawa,matubara}@el.itc.nagoya-u.ac.jp
‡National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{uchimioto,isahara}@nict.go.jp

Abstract

Our research organization has been constructing a large scale database named SHACHI by collecting detailed meta information on language resources (LRs) in Asia and Western countries. The metadata database contains more than 2,000 compiled LRs such as corpora, dictionaries, thesauruses and lexicons, forming a large scale metadata of LRs archive. Its metadata, an extended version of OLAC metadata set conforming to Dublin Core, have been collected semi-automatically. This paper explains the design and the structure of the metadata database, as well as the realization of the catalogue search tool.

1 Introduction

The construction of LRs such as corpora, dictionaries, thesauruses, etc., has boomed for years throughout the world in its aim of encouraging research and development in the main media of spoken and written languages, and its importance has also been widely recognized. Of the organizations willing to store and distribute LRs, there exist some consortia fulfilling their function such as LDC[1], ELRA[2], CLARIN[3], and OLAC[4], in Western countries, and GSK[5] which does so mainly in Japan. However, those released LRs are scarcely connected with each other because of the difference between written and spoken language as well as the difference between languages such as Japanese, English, and Chinese (OLAC User Guide, 2008).

This situation makes it difficult for researchers and users to find LRs which are useful for their researches. In the meantime, by connecting systematically existing various LRs with Wrapper Program, the attempt to realize multilingual translation services has already begun (Ishida et al, 2008, Hayashi et al, 2008). Moreover, since language information tags given to those LRs and their data formats are multifarious, each LR is operated individually. As LR development generally entails enormous cost, it is highly desirable that the research efficiency be enhanced by systematically combining those existing LRs altogether and extending them, which will encourage an efficient development of unprecedented LRs.

Our research organization has been constructing a large scale metadata database named SHACHI[6] by collecting detailed meta information on LRs in Western and Asian countries. This research project aims to extensively collect metadata such as tag sets, formats, and usage information about researches on those LRs. and recorded contents of LRs existing at home and abroad and store them systematically. Meanwhile, we have already developed a search system of LRs by the use of meta information and are attempting the experiment of widely providing meta information on our stored

[1]LDC:Linguistic Data Consortium, http://www.ldc.upenn.edu/

[2]ELRA: European LRs Association

[3]CLARIN: Common Language Resources and Technologies Infrastructure, http://www.ilsp.gr/clarin_eng.html

[4]OLAC: Open Language Archives Community, http://www.language-archives.org/

[5]GSK: Gengo Shigen Kyokai; Language Resource Association, http://www.gsk.or.jp/

[6]SHACHI: Metadata Database of Language Resources-SHACHI, Shachi means "orca" in English.

Coling 2008: Companion volume – Posters and Demonstrations, pages 119–122
Manchester, August 2008

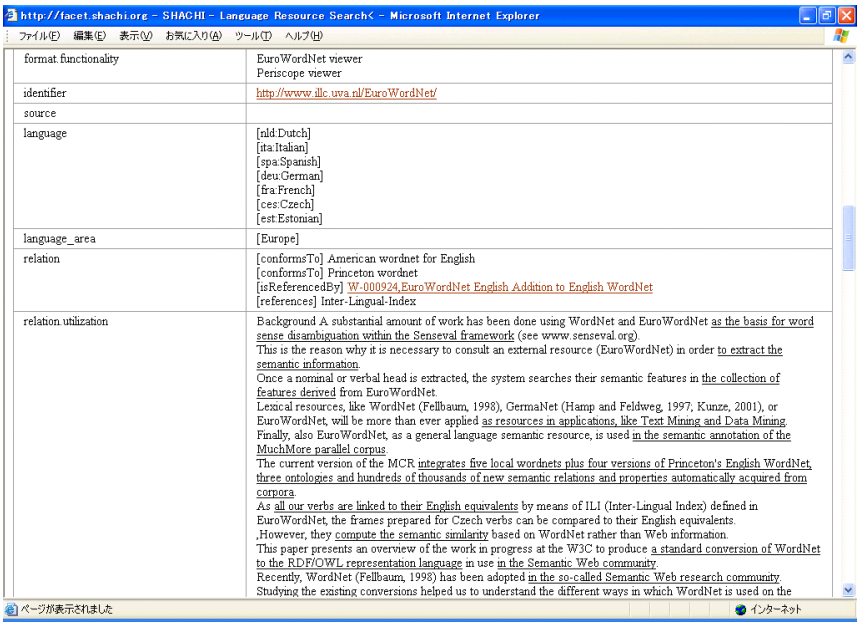

Figure 1: A sample page of SHACHI catalogue (ex. Euro WordNet)

LRs to those from researchers to common users. This metadata database has been now open to the public in the Web and allows every Internet user to access it for the search and read information of LRs at will.

2 Purpose of Metadata Database Construction

The purpose of the construction of the database is the following fivefold.

1. **To store language resource metadata:** SHACHI semi-manually collects detailed metadata of language resources and constructs their detailed catalogues. Figure 1 shows a sample page of a LR catalogue stored in SHACHI (ex. Euro WordNet). The catalogue provides more detailed meta information than other LR consortia do.

2. **To systematize language resource metadata:** Language resource ontology is tentatively constructed by classifying types of language resources (in this paper, it is called "ontology"). Figure 2 shows an example of its ontology. At the moment, it is under investigation what is the most useful and functional ontology for users by developing some ontologies such as human-made ontology, semi-automatically produced ontology, and automatically produced ontology.

3. **To make each language resource related to each other:** The detailed metadata enabled us to describe characteristics of each language resource and to expectably specify relationships among language resources. Figure 3 shows a part of the SHACHI search screen. It shows language resources found as a search result, the references to which these language resources conform as well as other language resources whose formats are common to theirs.

4. **To statistically investigate language resources:** By statistically analyzing the metadata, users are able to grasp what kinds of language resources exist in different part of the world and to understand current tendencies of language resources which have been available to the public.

5. **To promote the distribution of language resources:** Since this metadata database enables users to easily gain access to language resources in accordance with their needs, owing to fully equipped search functions, SHACHI will be able to support an effective use and an efficient development of language resources.

Some 2,000 resources of metadata have already been collected in the database so far and they will be enlarged by a further 3,000. To that end, it is

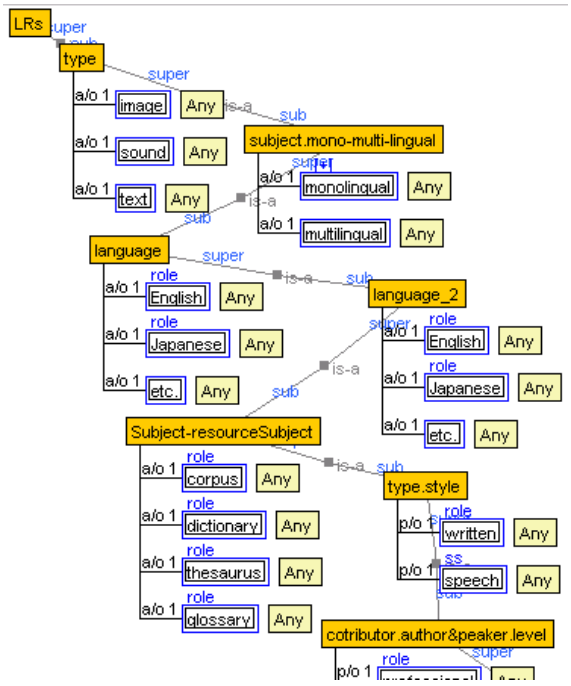

Figure 2: Automatically produced ontology

indispensable for us to work in cooperation with language resource consortia at home and abroad and to take the initiative in contributing to Asian language resources.

3 SHACHI Metadata Set

3.1 Policy for Collecting Metadata

The LRs which our metadata database stores should satisfy the following conditions:

- Those resources should be stored in a digital format.
- Those resources should be one of the following: corpus, dictionary, thesaurus, or lexicon. (Numeric data are not considered to be the subject of collection for SHACHI.)
- Those resources should be collected from English websites and its data must be open to the public.
- Those resources should be created by research institutions, researchers, or business organizations. (Developed tools such as facet search.)

LRs metadata database SHACHI covers meta information provided by LR consortia such as ELRA, LDC, and OLAC whose more detailed metadata are fed into the database by semi-automatic means of importing.

3.2 Extensions of Metadata Element

Since users sometimes search for LRs without a clear objective, it is necessary for language resource providers to construct language resource ontology. This database conforms to the OLAC metadata set which is based on 15 kinds of fundamental elements of Dublin Core[7] and constitutes an extended vision of OLAC with 19 newly added metadata elements which were judged to be indispensable for describing characteristics of LRs. SHACHI provides usage information about how and in which situation language resource researchers utilized each language resource, which is also important for users. The usage information about LRs is automatically retrieved from academic article databases (Kozawa et al, 2008). (See "Utilization" in Figure1).

3.3 Systematic Storage of LRs

Clear description of the relations among LRs can be applied to the efficient development of LRs and search tools for common users of database. Figure 2 shows ontology generated through automatic means, based on language resource metadata stored in SHACHI. We first surveyed the frequency of possible values of metadata element choicesand generated the ontology by hierarchicalizing meta elements of our meta categories. While ontology can be constructed in various ways from different standpoints, our ontology is particularly designed for users to enable to find them efficiently by following the hierarchical classes of our ontology.

4 Search Tools for Providing Users-Oriented Information

Figure 3 shows a screen image of a search result through SHACHI. This section discusses three search functions provided in SHACHI.

4.1 Three Types of Search Functions

For the purpose of facilitating users of this metadata database to find their intended language resource catalogues, SHACHI provides three search functions:

1. **Keyword search function:** This tool is suitable for users who have clear images to search

[7]Dublin Core Metadata Initiative, http://dublincore.org/

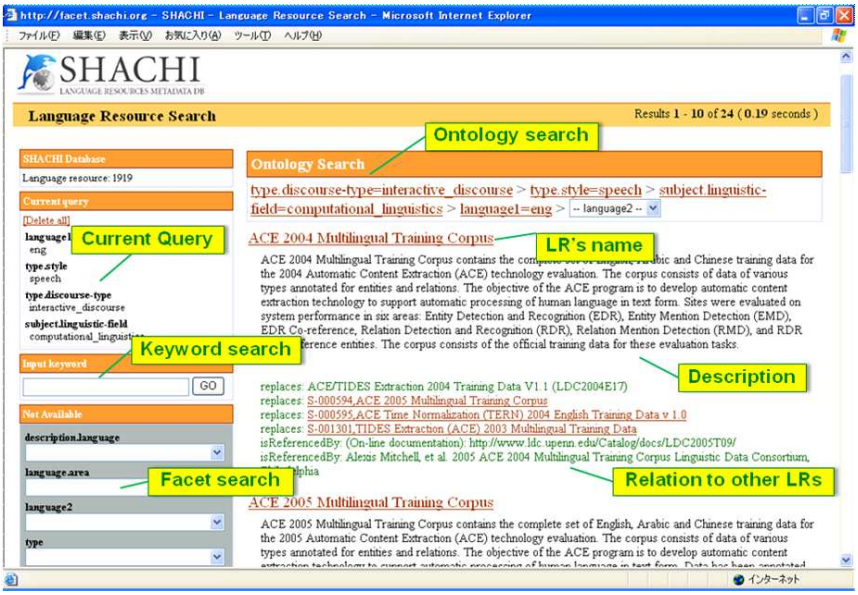

Figure 3: Catalogue search tool

for specified LRs and a technical knowledge of language processing. It allows them to input keywords as they want and to search all words stored in SHACHI metadata archive.

2. **Facet search function:** This tool is suitable for users who have a vague idea of what kind of LR they want. It is equipped with a choice of 15 kinds of metadata elements selected from the SHACHI metadata set. The users narrow down the target LRs one by one in order to find the intended one. For example, with one click on “age”, three choices such as “Childrenfs utterance?”, “Adultsf utterance?” and “Both are OK?” will be shown.

3. **Ontology search function:** This tool was developed by adopting the idea acquired by systematizing LRs registered in SHACHI. When using the ontology search function, users find the intended LRs by following the vertical relationship of the ontology. It was ascertained that ontology search function tool had the merit of enabling users to discover LRs that have not been ever found by keyword search and facet search functions.

5 Conclusion

In this paper, we reported on the design of SHACHI, a metadata database of LRs now being developed, the expansion and construction of metadata for it, and an actualization of a search function. At present, it contains approximately 2,000 pieces of meta information on LRs such as corpora, dictionaries and thesauruses. One of SHACHIfs characteristic features is that with a collection of tag sets, format samples, and usage information on LRs which is automatically retrieved from scholarly papers given to LRs. From now on the SHACHI project is intended to promote cooperation among other LRs consortia abroad as well as in Japan and to take the initiative in contributing to the development of LRs in Asia.

References

Ishida,T., Nadamoto, A., Murakami,Y., Inaba, R. et al. 2008. A Non-Profit Operation Model for the Language Grid, In proceedings of the 1st International Conference on Global Interoperability for language Resources, pp.114-121.

Kozawa, S., Tohyama, H., Uchimoto, K., Matsubara, S., and Isahara. H. 2008. Automatic Acquisition of Usage Infor-mation for Language Resources, In proceedings of the 6th edition of the Language Resources and Evaluation Conference.

OLAC (Open Language Archives Communi-ty), 2008. Searching of OLAC Metadata: User Guide, http://www.language-archives.org/tools/search/searchDoc.html

Yoshihito Hayashi, Thierry Declerck, Paul Buitelaar, Monica Monachini. 2008. Ontologies for a Global Language Infrastructure, In proceedings of the 1st International Conference on Global Interoperability for language Resources, pp.105-112.