

Infrastructure for Language Service Composition

Yohei Murakami

Language Grid Project,
National Institute of
Information and
Communications Technology
yohei@nict.go.jp

Toru Ishida

Language Grid Project,
National Institute of
Information and
Communications Technology/
Kyoto University
ishida@i.kyoto-u.ac.jp

Takao Nakaguchi

NTT Advanced Technology
Corporation
takao.nakaguchi@ntt-at.co.jp

Abstract

Although English has become the standard language in various areas, most people do not use it in local activities. To increase the mutual understanding of different cultures with different languages, it is essential to build a language infrastructure on top of the Internet that improves the accessibility and usability of existing online language services so that users can create new cross-language services for their communities. To realize this infrastructure, this paper proposes the language grid. The language grid consists of the "horizontal language grid," which connects the standard languages of nations, and "vertical language grid," which combines the language services generated by communities. This approach can facilitate intercultural collaboration through the Internet, such as international online collaborative learning.

1. Introduction

2001.9.11 impacted the world. We remember there was a clear conflict in public opinion in western countries. While 77% of those interviewed in France opposed military intervention against Iraq (2003.1.9 Le Figaro), 63% of the U.S. population were proud of the U.S. role in the war (2003.3.22 CBS News). Conflicts in governmental policies are common, but conflicts in public opinion between western countries of this scale have not been observed before. Though we all share information on the Web, similar conflicts arose recently in East Asia. While about 90 percent of Chinese polled blamed Japan for the World War II responsibility, more than half of Japanese polled said it was hard to tell who bore responsibility (2005.8.24 Genron NPO and Peking University). According to Global Reach, the ratio of English speaking people online has decreased to 35.2% in

2004. To increase mutual understanding between different cultures with different languages, it is essential to build a language infrastructure on top of the Internet.

Motivated by the above goal, we conducted Intercultural Collaboration Experiments in 2002 (ICE2002) with Chinese, Korean and Malaysian colleagues [7]. We thought that machine translation would be useful in facilitating intercultural cooperation in advancing a software project. We gathered machine translators to cover five languages: Chinese, Japanese, Korean, Malay, and English. More than forty students and faculty members from five universities in four countries joined this experiment. The goal was to develop open source software using the participants' first language. The experiment started in April 2002 and ended in December 2002, and a total of 31,000 messages were collected.

2. Language grid

We provided translation services covering five languages for ICE2002. From this experiment, we found that language services are often not accessible, because of intellectual property rights and cost. We tend to think that effective language infrastructures have been already developed, since we can use machine translations to view Web pages. However, if you try to create new services by combining existing language services, you will soon be forced to face the realities: the language services available come with different contracts and prices. Contracts can be complex because of the concern over intellectual property rights. Prices can be high, and no explanation is available. Furthermore, language services are often not usable, because of non-standardized interfaces. Users have to develop different wrappers for different language services. There is no quality assurance for machine translators. Users have to estimate their quality by themselves. Existing services are

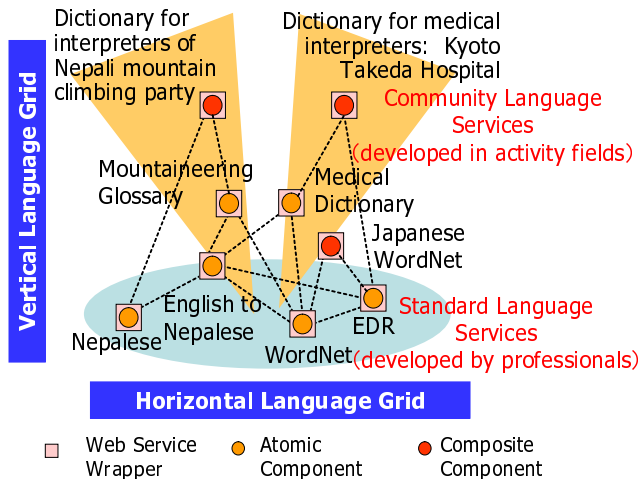


Figure 1. Language grid.

often not customizable: machine translators seldom allow users to modify them; it is hard to add new words to user dictionaries.

To increase the accessibility and usability of language services, we proposed the language grid, which treats existing language services as atomic components and enables users to create new language services by combining the appropriate components [4]. The feature of the language grid is to integrate language services, which is different from EuroWordNet [8] integrating European lexical data. Note that language services consist of language resources including dictionaries, thesauri and corpuses, and language processing functions including morphological analysis, translation, and paraphrasing.

The language grid has two different goals. One is to connect existing online language services that cover standard national languages. Those services are created often by linguistic professionals with the support of their governments. Typical examples include online dictionaries and translation services. Another goal is to assist users to create new language services, which are often related to intercultural activities in their local community. Consequently, the language grid consists of two different types of service networks as shown in Figure 1.

The horizontal language grid combines existing language services using semantic Web technology. For example, by connecting WordNet to an English-Japanese dictionary, we can create WordNet with a Japanese interface. The horizontal language grid benefits a wide range of users by providing standard language services. The vertical language grid, on the other hand, layers community language services on the horizontal language grid to support intercultural activities. The language service ontology is introduced to represent entries of language resources and processing

functions. The ontology also enables users to easily extend default service interfaces to define community-oriented services [2]. Suppose a nonprofit organization has its own parallel texts to support foreigners in a specific affiliated hospital. The organization can combine standard parallel texts created by medical doctors and its own resources by using the language grid.

3. Design of language grid

The key to realize the language grid is to coordinate language services developed by various providers and developers. Since the language services are implemented as Web services with different interfaces in different programming languages, we have to coordinate the language services using some approach that is implementation independent. One solution to the interoperability of applications, Web service technology, has gain attention recently. Web service is a generic term of XML-based technology to coordinate applications by some platform independent technology; examples are SOAP (Simple Object Access Protocol) and WSDL (Web Service Description Language). In addition, BPEL4WS (Business Process Execution Language for Web Services)[1] was proposed as a language to describe workflows for coordinating Web services. In this section, we explain the system design of the language grid based on Web service technology.

3.1. Use case of language grid

Users of the language grid are divided into two types of users: *language service providers* who publish their language services on the language grid, and *language service users* who utilize language services deployed on the language grid. Language service providers implement their

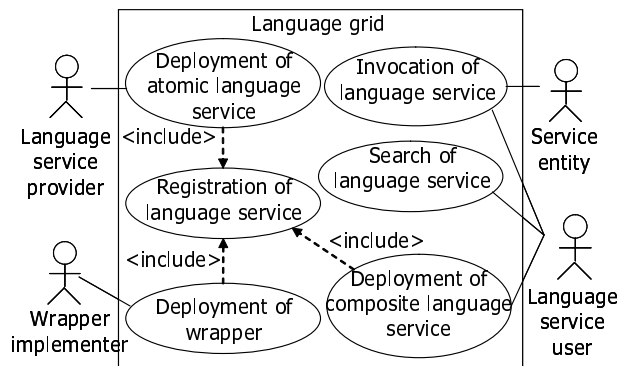


Figure 2. Use cases of language grid.

own language service as a Web service with a standard interface, deploy them on the language grid, and register their WSDL description and profile on the language grid. Meanwhile, language service users locate the services they need on the language grid and then invoke them. At the same time, they can construct a new composite language service by applying the workflow, and then deploy the services. Also, even when a language service is implemented without the standard interface, it can be integrated into the language grid by a third party (the wrapper implementer), who deploys a wrapper to adjust its interface to the standard interface. In this case, the service entity of the language service is located outside the language grid and is invoked by the wrapper. Figure 2 shows these use cases of the language grid.

3.2. Language grid network

The grid is defined as a system that coordinates resources that are not subject to centralized control, and delivers non-trivial qualities of service. To equip the language grid with such a function, we construct the language grid with two different types of server nodes: one is called the *language grid service node* and the other is called the *language grid core node*. Figure 3 illustrates how these two types of nodes are linked.

The *language grid service node* provides only language services. On the language grid service node, not only service entities but also wrappers to standardize interfaces of external language service entities are deployed. On the other hand, the *language grid core node* manages registration information of language services and coordinates language services.

In Figure 3, the arrows between nodes show the direction

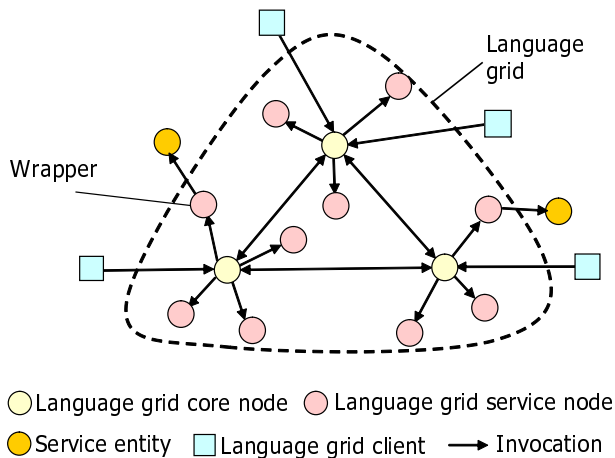


Figure 3. Relation between server nodes.

of service invocation. At first, a client tool for the language grid sends a message of service invocation to language grid core node. Upon receiving the message, the language grid core node invokes other language services according to the workflow of the invoked service. The invoked language services can be other composite language services provided by language grid core nodes as well as atomic language services provided by language grid service nodes. Moreover, when wrappers on language grid service nodes are invoked, they also invoke service entities outside of the language grid.

3.3. System architecture of language grid

To construct a network consisting of two types of language grid nodes, we created the system architecture shown in Figure 4. The system consists of three components; language grid core node, language grid service node, and language grid client.

3.3.1. Language grid core node Language grid core node consists of three functions, "Language service registration/deployment function," "Language service search function," and "Composite language service execution function," and two repositories, "Language service information repository," and "Workflow repository". All functions are provided as Web services so that users can employ them using standard, open, general-purpose protocols and interfaces.

Language service registration/deployment function

This function stores the WSDL description and the profile of a language service from a language service provider in language service information repository. A language service whose WSDL description and profile are stored in the repository is available to language service users. Workflows described by language service users can be deployed on the language grid as composite language services.

Language service search function

This function locates language services that will satisfy the user's need among those registered on the language grid. Specifically, it provides search by languages processable by the language services or types of language services. If there is a language service that matches the user's demand, the function returns the WSDL description and profile of the language service to the user.

Composite language service execution function

This function executes a composite language service registered on the language grid according to a request

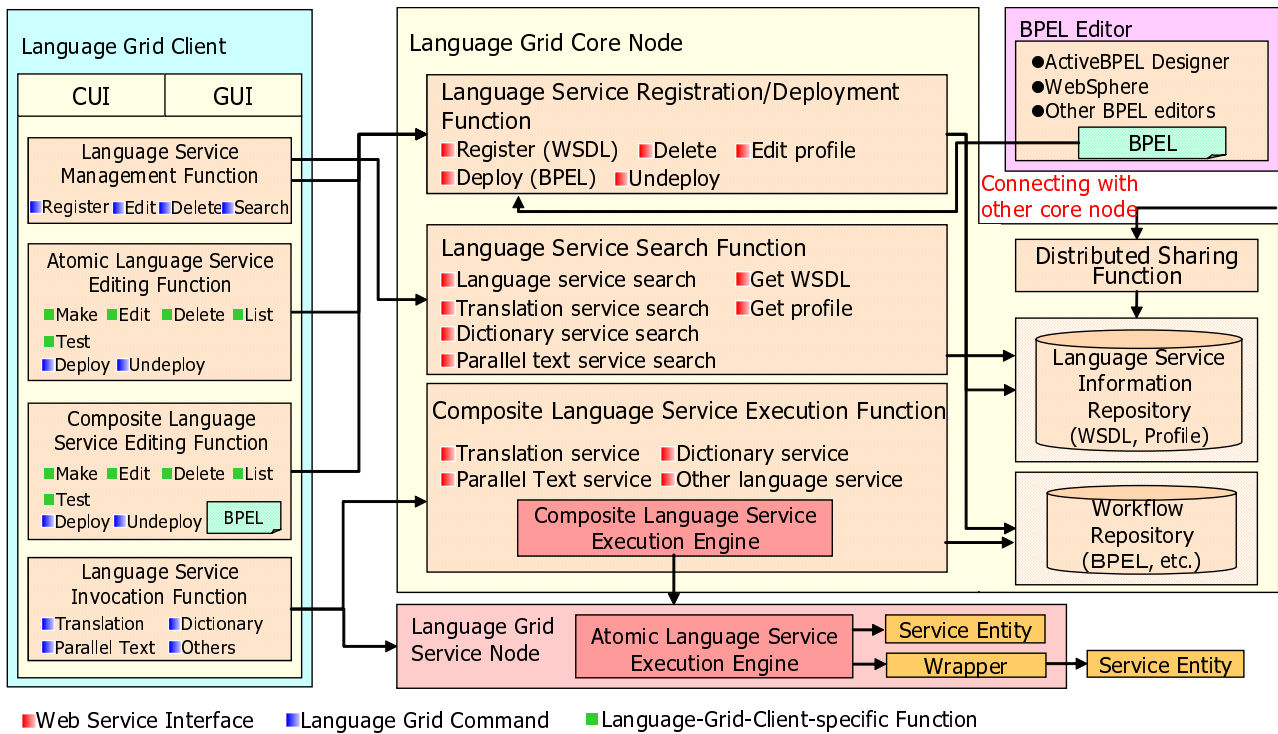


Figure 4. System architecture of language grid.

message from a language service user. At first, it retrieves a BPEL file to accomplish the request of the user from the workflow repository and then executes it using a workflow execution engine for composite language services. The composite language service execution engine invokes atomic language services on the language grid service nodes or other composite language services on the language grid core nodes following the workflow.

Language service information repository

This repository manages the WSDL descriptions and profiles of language services registered by language service registration/deployment function. The information within the repository is accessed by the language service search function in order to find the language services the user needs.

Workflow repository

This repository manages workflows (BPEL description) deployed by language service registration/deployment function. The workflows are executed by the workflow execution engine by composite language service execution function for a composite language service.

3.3.2. Language grid service node Each language grid service node has an execution engine to invoke atomic language services and a function to deploy service entities of atomic language services and wrappers of external service entities. Service entities and wrappers are deployed as Web services with standard interfaces. Since the wrappers do not have service entities on the language grid service nodes, they invoke existing service entities outside the language grid. The wrappers format the result into an output message defined in the standard interface and return it to the service invoker.

3.3.3. Language grid client The language grid client is a tool to enable language service users to easily use language grid core nodes. It provides both CUI and GUI based user interfaces. Each function provided by the language grid core node has a matching language grid client function. Thus all the functions of the language core node can be controlled by the language grid client.

Furthermore, the language grid client has a function to edit workflows for the coordination of multiple language services. Workflows composed by this function can be deployed on the language grid core node with the language grid client. In addition, since we adopt BPEL4WS as the workflow description language, which is becoming a de facto standard in the web service composition domain, the

coordination workflow of a language service described by another BPEL editor can be deployed on the language grid core node.

4. Application: tailored translation

The increase in the frequency of intercultural collaboration through the Internet, such as online collaborative learning and open source development, requires a variety of community language services. One example of language services that support communities is the Web-based collaborative translation environment, "Yakushite Net (<http://www.yakushite.net>) [6]." "Yakushite Net" enables people with deep knowledge of a particular domain to collaborate in enhancing the specialized dictionaries for online machine translation, and thus realize more accurate translations specific to a community. However, the mechanism forces us to customize the machine translation engine in order to integrate the dictionary created by a community. As a result, even when a community has another machine translation with higher accuracy, it cannot use the machine translation without costing the community much effort to customize it. Therefore, we require a method that can customize workflows to coordinate language services including machine translations rather than the machine translation engine itself. This method enables us to replace a language service with another language service whose interface is the same; this improves translation scalability.

4.1. Tailored translation workflow

In order to create a workflow of a tailored translation like "Yakushite Net," we first have to wrap existing morphological analyses and machine translations in our standard interface, and deploy them on a language grid service node. Meanwhile, we must implement a bilingual dictionary of technical terms as a Web service with our standard interface and deploy it on a language grid service node in the same manner. Next, we register the WSDL description and profile of the deployed service. This makes the service available to language service users.

Figure 5 illustrates the workflow of tailored translation; it consists of the morphological analysis service, the bilingual dictionary of technical terms service, and the machine translation service. In this workflow, the composite service divides the input sentence into a set of morphemes using the morphological analysis service. Next, it extracts technical terms from the set of morphemes and creates a set of technical terms included in the input sentence. Next, it obtains accurate translations of the technical terms from the bilingual dictionary. In the same way, it also obtains intermediate codes of the technical terms, strings that have little influence on machine translation. After that, it replaces

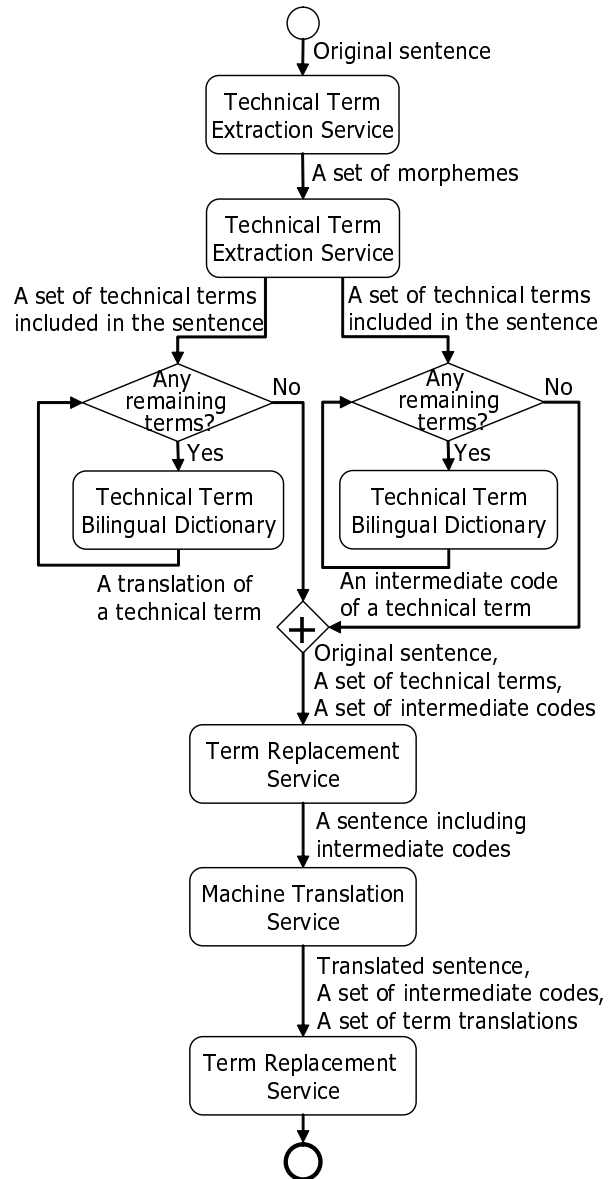


Figure 5. Workflow of tailored translation.

the technical terms in the input sentence with the intermediate codes and translates the sentence including intermediate codes using the machine translation service. Finally, it replaces the intermediate codes in the translated sentence with translations of the technical terms, formats the result into an output message defined in the standard translation interface, and returns it to the service invoker.

Since this workflow is an abstract description, it can generate various types of tailored translations by specifying concrete morphological analysis services, bilingual dictionaries of technical terms, and machine translations. In fact, using this workflow, we have created an instance of

the tailored translation service for supporting participants in NDYS (Natural Disaster Youth Summit) project organized by NPO iEARN, and facilitators of NPO Pangaea's activity.

5. Conclusion

To support intercultural collaboration, this paper proposes the *language grid* to increase the *accessibility* and *usability* of language services. As is clear, field study is essential in researching the language grid and understanding how language services should be created in local communities. In the future, by describing the semantics of language services in OWL-S [5], we expect that the language grid can also work as a language infrastructure for a translation agent that supports intercultural communication [3].

Acknowledgement

This research was done with K. Okamoto, Y. Fujihara, K. Fujii, A. B. Hassine, Y. Hayashi, R. Hishiyama, R. Inaba, C. Kita, S. Matsubara, T. Masuura, Y. Mori, A. Nadamoto, H. Nakanishi, Y. Naya, A. Shigeno, T. Shigenobu, T. Takasaki, and T. Yoshino. Translation services for this research were provided by Kodensha Co., Ltd and Cross Language Inc, and morphological analysis services by Kyoto University, Kookmin University, Chinese Academy of Science, and University Stuttgart.

References

- [1] T. Andrews, F. Curbera, H. Dolakia, J. Goland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte, I. Trickovic, and S. Weeravarana. *Business Process Execution Language for Web Services*, 2003.
- [2] Y. Hayashi and T. Ishida. A dictionary model for unifying machine readable dictionaries and computational concept lexicons. In *Proc. of the International Conference on Language Resources and Evaluation (LREC06)*, 2006.
- [3] T. Ishida. Communicating culture. *IEEE Intelligent Systems*, 21(3):62–63, 2006.
- [4] T. Ishida. Language grid: An infrastructure for intercultural collaboration. In *Proc. of IEEE/IPSJ Symposium on Applications and the Internet (SAINT06)*, pages 96–100, 2006.
- [5] D. Martin, M. Paolucci, S. McIlraith, M. Burstein, D. McDermott, D. McGuinness, B. Parsia, T. Payne, M. Sabou, M. Solanki, N. Srinivasan, and K. Sycara. Bringing semantics to web services: The owl-s approach. In *International Workshop on Semantic Web Services and Web Process Composition*, 2004.
- [6] T. Murata, M. Kitamura, T. Fukui, and T. Sukehiro. Implementation of collaborative translation environment: Yakushitenet. In *Proc. of the 9th Machine Translation Summit System Presentation*, pages 479–482, 2003.
- [7] S. Nomura, T. Ishida, N. Yamashita, M. Yasuoka, and K. Funakoshi. Open source software development with your mother language: Intercultural collaboration experiment. In *Proc. of the International Conference on Human-Computer Interaction (HCI03)*, pages 1163–1167, 2003.
- [8] P. Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173, 2004.