Cognitive Technologies

Toru Ishida *Editor*

Service-Oriented Collective Intelligence for Language Resource Interoperability

Cognitive Technologies

Toru Ishida
Editor

The Language Grid

Service-Oriented Collective Intelligence for Language Resource Interoperability

Springer

Editor
Prof. Toru Ishida
Department of Social Informatics
Kyoto University
Yoshida-Honmachi
606-8501 Kyoto
Japan
ishida@i.kyoto-u.ac.jp

Managing Editors
Prof. Dov M. Gabbay
Augustus De Morgan Professor of Logic
Department of Computer Science
King's College London
Strand, London WC2R 2LS, UK

Prof. Dr. Jörg Siekmann
Forschungsbereich Deduktions- und
Multiagentensysteme, DFKI
Stuhlsatzenweg 3, Geb. 43
66123 Saarbrücken, Germany

Cognitive Technologies ISSN 1611-2482
ISBN 978-3-642-21177-5 e-ISBN 978-3-642-21178-2
DOI 10.1007/978-3-642-21178-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011934512

ACM Codes: I.2.7, J.5, H.3

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The idea of *the Language Grid* was born from the long term research initiative of *intercultural collaboration* in Kyoto University as follows. The concept of intercultural collaboration was invented after 9.11 in 2001 to establish the practical research goal of cross-cultural study: measuring the progress made in this study is feasible if we focus on collaborative work rather than general communication. We targeted machine translation as a key technology for intercultural collaboration, because only machine translators can overcome language barriers.

We took the approach of creating innovation from studies in the field. In 2002, we conducted a one year experiment called the *Intercultural Collaboration Experiment (ICE2002)* with Chinese, Korean, and Malaysian colleagues. They used their mother tongues to jointly develop open source software. After trying to collect and customize machine translators to cover five languages, we realized that the Internet needed a language infrastructure. Though there are language resource associations scattered throughout the world, their missions are to provide language resources to researchers, not to end users. We need an infrastructure for end users that provides unhindered access to language resources worldwide and that allows them to be combined to create customized multilingual environments.

In 2005, we designed a service-oriented language infrastructure, and submitted a research proposal to the National Institute of Information and Communications Technology. Our idea is to shift *from language resources to language services*. The research goal is not to collect language resources but to interconnect them as Web services. Since language is used everywhere in our daily life, delivering general technologies is not enough to cope with language barriers: we need to encourage end users to create their own customized multilingual environments to cover the situations they face. We think that the language barriers created by billions of people will be overcome by the very same people. The research proposal was accepted and the Language Grid Project commenced in 2006.

This book includes eighteen chapters in six parts that summarize various research results and associated development activities on the Language Grid. The first part describes the framework of the Language Grid, *service-oriented collective intelligence*, to bridge service providers, service users, and service grid operators. Two kinds of software are newly introduced: the *service grid server software* and the *Language Grid Toolbox*. The former is able to form any kind of service grid such as language services, e-learning services, and agricultural knowledge services. The Language Grid has been created based on this software and various language services. The latter is client software to create customized multilingual environments for end users. Both source codes are available with open source licenses. The second part provides technologies for *service workflows* that compose atomic language services. Since there can be different language services with the same functionalities, *horizontal service composition* to select the optimal atomic services for a given workflow is proposed based on constraint optimization algorithms. To this end, *language service ontology* has been studied to specify the se-

mantics of various language services. To monitor and control the execution of composed workflows, the notion of *service supervision* was introduced.

The third part reports research work and activities on sharing and using language resources. Pivot translation is often applied, for example, when Asian languages are translated into European languages via English. Context-aware supervision has been proposed to prevent the decrease in translation accuracy that has, up to now, been inevitable with cascading translation services. Researchers whose background is not natural language processing also started using language resources: cultural differences in pictograms drawn by kids were analyzed by using a concept dictionary. NPO staff developed a system by which volunteer interpreters could collectively create questions and answers in different languages to support foreign patients at hospital receptions. The fourth part provides various applications of language services as applicable to intercultural collaboration. International NPOs have been using translation services for daily communication among their volunteer members worldwide. The accuracy of translation has been improved by developing their own dictionaries to be used in conjunction with translators. Researchers created a language-barrier-free room in the Second Life virtual space to naturally observe informal communication using machine translators. In parallel, controlled experiments have been conducted to understand the *inconsistency*, *asymmetry* and *intransitivity* of machine translations.

The fifth part collects reports on applying the Language Grid for translation activities including localization of industrial documents and Wikipedia articles. Protocols for *collaborative translation* have been studied to guarantee the effectiveness of group activities for translation. The sixth part illustrates how the Language Grid can be connected to other service grids. The Language Grid has already been combined with *Heart of Gold*, which was developed by DFKI in Germany for pipelining language processing software. The Language Grid has been also connected with smart classroom services in Tsinghua University in China. Furthermore, the two operation centers of the Language Grid in Kyoto University and NECTEC have been federated for joint operation.

We hope this book will strongly support and encourage researchers who are willing to utilize various language resources worldwide to create customized multilingual environments to overcome local language barriers. We are grateful to the many people who have worked on, collaborated with, and supported the Language Grid Project. This project will continue to guarantee the free exchange of ideas in different languages to prevent serious cultural conflicts in the future.

Toru Ishida
March, 2011 in Kyoto

Contents

Part I Language Grid Framework

Chapter 1 The Language Grid: Service-Oriented Approach to Sharing Language Resources ……..…….……. 3
Toru Ishida, Yohei Murakami, and Donghui Lin

Chapter 2 Service Grid Architecture …….…….……..… 19
Yohei Murakami, Donghui Lin, Masahiro Tanaka, Takao Nakaguchi, and Toru Ishida

Chapter 3 Intercultural Collaboration Tools Based on the Language Grid .....…. 35
Masahiro Tanaka, Rieko Inaba, Akiyo Nadamoto, and Tomohiro Shigenobu

Part II Composing Language Services

Chapter 4 Horizontal Service Composition for Language Services …… 53
Ahlem Ben Hassine, Shigeo Matsubara, and Toru Ishida

Chapter 5 Service Supervision for Runtime Service Management …….…..... 69
Masahiro Tanaka, Toru Ishida, and Yohei Murakami

Chapter 6 Language Service Ontology …….…….…….… 85
Yoshihiko Hayashi, Thierry Declerck, Nicoletta Calzolari, Monica Monachini, Claudia Soria, and Paul Buitelaar

Part III Language Grid for Using Language Resources

Chapter 7 Cascading Translation Services …….…… 103
Rie Tanaka, Yohei Murakami, and Toru Ishida

Chapter 8 Sharing Multilingual Resources to Support Hospital Receptions ….. 119
Mai Miyabe, Takashi Yoshino, and Aguri Shigeno

Chapter 9 Exploring Cultural Differences in Pictogram Interpretations …..…. 133
Heeryon Cho and Toru Ishida

Part IV Language Grid for Communication

Chapter 10 Intercultural Community Development for Kids around the World .151
Toshiyuki Takasaki, Yumiko Mori, and Alvin W. Yeo

Chapter 11 Language-Barrier-Free Room for Second Life …….……...…..... 167
Takashi Yoshino and Katsuya Ikenobu

Chapter 12 Conversational Grounding in Machine Translation Mediated Communication ……..… 183
Naomi Yamashita and Toru Ishida

Part V Language Grid for Translation

Chapter 13 Humans in the Loop of Localization Processes ……..…. 201
Donghui Lin

Chapter 14 Collaborative Translation Protocols ……...…. 215
Daisuke Morita and Toru Ishida

Chapter 15 Multi-Language Discussion Platform for Wikipedia Translation ... 231
Ari Hautasaari, Toshiyuki Takasaki, Takao Nakaguchi, Jun Koyama, Yohei Murakami, and Toru Ishida

Part VI Towards Federation of Service Grids

Chapter 16 Pipelining Software and Services for Language Processing ……. 247
Arif Bramantoro, Ulrich Schäfer, and Toru Ishida

Chapter 17 Integrating Smart Classroom and Language Services …… 263
Yue Suo, Yuanchun Shi, and Toru Ishida

Chapter 18 Federated Operation Model for Service Grids …… 279
Toru Ishida, Yohei Murakami, Eri Tsunokawa, Yoko Kubota, and Virach Sornlertlamvanich

Biography of Authors 299

Part I

Language Grid Framework

Chapter 1
The Language Grid: Service-Oriented Approach to Sharing Language Resources

Toru Ishida[1], Yohei Murakami[2], and Donghui Lin[2]

1 Department of Social Informatics, Kyoto University, Yoshida Honmachi, Sakyoku, Kyoto 606-8501 Japan, e-mail: ishida@i.kyoto-u.ac.jp

2 Language Grid Project, NICT, 3-5 Hikaridai, Seikacho, Sorakugun, Kyoto 619-0289, e-mail: {yohei, lindh}@nict.go.jp

Abstract Since various communities, which use multiple languages, now want to interact in daily life, tools that can effectively support multilingual communication are necessary. However, we often observe that the success of a multilingual tool in one situation does not guarantee its success in another. To develop a multilingual environment that can handle various situations in various communities, existing language resources (dictionaries, parallel texts, part-of-speech taggers, machine translators, etc.) should be easily shared and customized. Therefore, we designed our proposal, the Language Grid, as service-oriented collective intelligence; it allows users to freely create language services from existing language resources and combine those language services to develop new services to meet their own requirements. This chapter explains the design concept and service architecture of the Language Grid, and the approach of user involvement in the collective intelligence activities.

1.1 Introduction

Though the Internet allows people to be linked together regardless of location, language remains the biggest barrier: only 35% of the Internet population speaks English (Paolillo et al. 2003). The remainder is divided between other European languages and Asian languages. In fact, it is not possible for anyone to learn the languages needed to access all possible information on the Internet. In particular, Asian people are not taught neighboring languages. Few Japanese understand Chinese or Korean and vice versa. People learn English to collaborate, but often cannot think in English: serious barriers to intercultural collaboration exist, because the collaboration often requires elaborating new ideas in the native language. As there is no simple way to solve this problem, it is necessary to combine different ideas. Teaching English is one way, but learning another's language and

T. Ishida (ed.), *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*, Cognitive Technologies, DOI 10.1007/978-3-642-21178-2_1,

respecting another's culture are also important. Since one cannot master all languages, the use of machine translation systems is a viable solution.

The above background drove us to conduct the *Intercultural Collaboration Experiment 2002 (ICE2002)* with Chinese, Korean and Malaysian colleagues (Nomura et al. 2003). We thought that machine translation would be useful in facilitating intercultural exchanges. We gathered machine translators to cover five languages: Chinese, Japanese, Korean, Malay and English. More than forty students and faculty members from five universities joined this experiment. The goal of the experiment was to develop open source software using the participants' first languages: Japanese participants used the Japanese language, Chinese participants used the Chinese language, and so on. The experiment started in April 2002 and ended in December 2002. During this experiment, the following problems were found in using language resources. Note that language resources include dictionaries, parallel texts, part-of-speech taggers, machine translators and so on.

- *Language resources are often not accessible* because of intellectual property rights and prices. We can now see many new language services on the Internet. We tend to think that effective language infrastructures have been developed, since we can use machine translations to view Web pages. However, if one tries to create new services by combining existing language resources, he/she is soon forced to face the realities: the language resources available come with different contracts and prices. Contracts tend to be complex because of concern over intellectual property rights. Explanations of the pricing structure are often incomplete or confusing even if the price is high.
- *Language resources are often not usable*, because of nonstandard interfaces and low service quality. For application interfaces, users have to develop different wrappers for different language resources. There is no quality assurance for language processing software including machine translators. Users have to estimate their quality of services, when selecting one. Moreover, language resources are often not customizable. Machine translators seldom allow users to modify them; it is hard to add new words to their dictionaries.

To increase the accessibility and usability of language resources, we proposed the *Language Grid* as *service-oriented collective intelligence*, i.e., it wraps existing language resources as atomic services and enables users to compose new services by combining atomic services. To realize the Language Grid, however, we must deal with the following issues.

- *Service architecture*: The service platform should allow users to create services and share them. Based on various atomic services with standard interfaces, an infrastructure for service composition should be provided. The service architecture should also allow users to develop application systems for supporting multilingual activities in their communities based on the provided language services.
- *User involvement*: Collective intelligence platforms can grow only through the voluntary efforts of users (Weiss 2005). The more users provide resources, the

more they can utilize the benefits of the resources. Therefore, it is necessary to encourage the participation of both users and communities.

Researchers in several organizations including Kyoto University and National Institute of Information and Communications Technology (NICT) started working on the Language Grid in April 2006 (Ishida 2006). This project is based on collaboration between industry, government, universities and non-profit organizations (NPO/ NGOs). The remaining parts of this chapter are organized as follows. First, Section 1.2 explains the necessity of shifting from language resources to language services. Section 1.3 shows the design concept and the service architecture, and Section 1.4 introduces how the Language Grid is operated for user involvement.

1.2 From Language Resources to Language Services

This section describes why the service-oriented approach is promising for sharing language resources. To illustrate this, let us look at what would happen in a Japanese school, where the number of Brazilian, Chinese and Korean students is increasing. We use machine translators in this example.

Suppose the teacher says "You have cleanup duty today" in Japanese, it means "It is your turn to clean the classroom today," and foreign students cannot understand this. Puzzled students are invited to a multilingual room in the school. Sitting in front of a computer connected to the Internet, the teacher types these words in Japanese on the screen: "You have cleanup duty today." Then the translation of this sentence appears: "今天是你负责打扫卫生" in Chinese, "오늘은 네가 청소 당번이야" in Korean, and "Hoje é seu plantão de linpeza" in Portuguese. "Aha!" say the kids with excited faces. One of them types in their language "I got it" and translation appears in Japanese on the screen.

Is it that simple to use machine translation? Several portal sites already offer translation services. Let's try to use them. First, enter "You have cleanup duty today" in Japanese and translate it into Korean. The sentence "오늘은 너가 청소 당번이야" appears on the screen. The Japanese teachers do not understand Korean, so they are not sure if the translation is correct. They use *back-translation* which translates the Korean translation into Japanese again. This yields "You should clean the classroom today!" It seems a little rude to hear, but may be acceptable, if accompanied with a smile. Let's translate it into Chinese in the same way. The Chinese sentence "今天你是扫除值日哟" appears on the screen. The Japanese teachers back-translate this Chinese sentence into Japanese and find the very strange sentence "Today, you remove something to do your duty." It seems the Japanese word "掃除当番," which means duty to clean the classroom, was not registered in the dictionary of this machine translator.

It appears that we need to customize machine translators with local dictionaries. For schoolteachers, it is necessary to compile a multilingual dictionary of words frequently used in schools. Suppose the available multilingual dictionaries

are adequate. To combine those local resources and machine translators, however, we need to negotiate with the companies that provide the machine translators, and make contracts with them. Let's assume all contracts are signed successfully. It is still not easy to combine machine translators and dictionaries, because the APIs and data formats are not standardized.

The service-oriented approach allows users to create and share standardized dictionary services while protecting the intellectual property rights of language resources. Fig. 1.1 shows how to create atomic language services from corresponding language resources. Data like multilingual dictionaries and parallel texts can be wrapped to create atomic language services to provide a translation of words or sentences. However, those atomic services do not have to be a simple retrieval function: a parallel text service can return the translation of a sentence that is similar to the input sentence. Wrapping software like machine translators is straightforward. Even human interpreters can be wrapped as translation services. Users do not have to distinguish machine from human translation services other than by their quality of services: machine translators can provide faster services while human interpreters return higher quality translations.

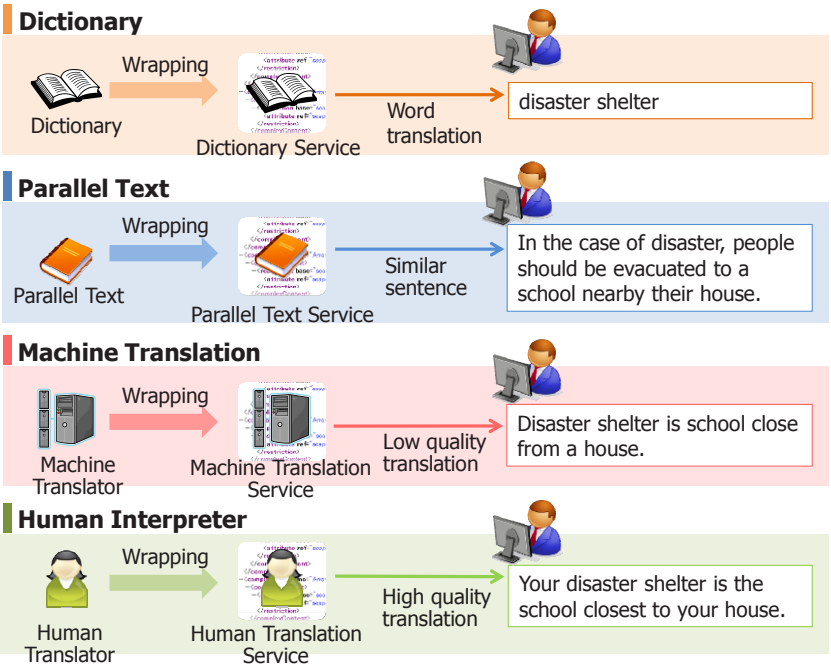

Fig. 1.1 Language service (atomic)

The next step is to compose atomic language services to create a new service. Fig. 1.2 illustrates the process of composing a variety of atomic language services for Japanese teachers to translate their announcements for Brazilian parents. To

translate Japanese sentences into Portuguese, we first need to cascade Japanese-English and English-Portuguese translators, because there is no available direct translator handling Japanese to Portuguese. To replace words output by machine translators with the words in multilingual dictionaries for schools, part-of-speech taggers are necessary to divide the input sentences into parts. We can train *example-based machine translators* with Japanese-Portuguese parallel texts. We then have different types of translators including example-based machine translators and will face the problem of determining which one is best: example-based machine translators can create high quality translation only when they trained with similar sentences. We may use back-translation, say Japanese-Portuguese-Japanese translation, to compare original and back-translated Japanese sentences, and select the translator that can produce back-translated sentences most similar to the original ones. If the quality of translation is still not enough for the Brazilian parents to understand, however, Japanese teachers may use human translation services to create an announcement in Portuguese.

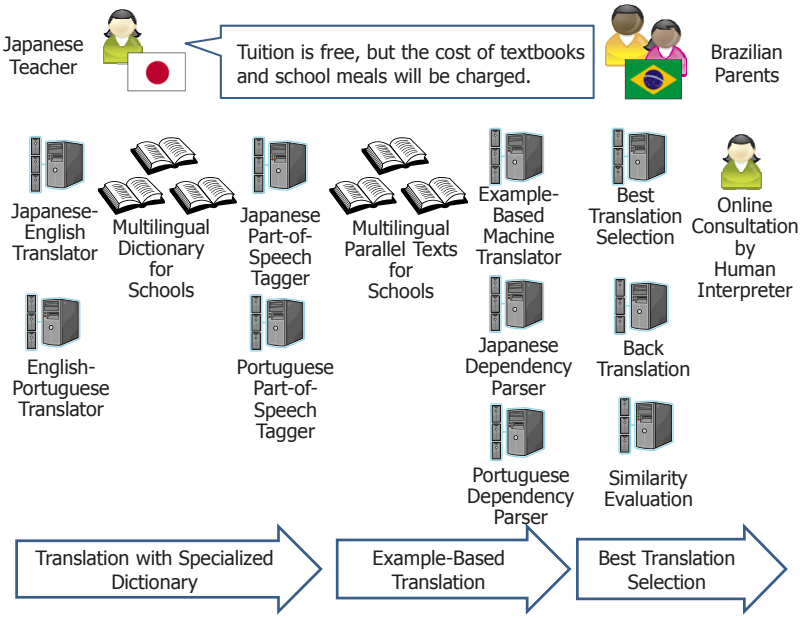

Fig. 1.2 Language service (composite)

By the way, part-of-speech taggers are often developed in research institutes or universities, and are provided only for research purposes. Their Web sites do not state that they can be used in schools, hospitals and so on. If an elementary school wants to use them, the school needs to ask those providers for permission by a letter or e-mail. One of the important roles of the Language Grid is to reduce such negotiation costs related to intellectual property rights.

1.3 The Language Grid

1.3.1 Design Concept of the Language Grid

As discussed in the previous sections, language resources already exist online. However, difficulties often arise when people try to use those language resources in their intercultural activities; complex contracts, intellectual property rights, and non-standard application interfaces make it difficult for users to create customized language services that support intercultural activities. To improve the accessibility and usability of existing language resources, we need to allow users to easily create new language services by combining existing ones. As shown in Fig. 1.3, the Language Grid should provide an environment where users can share language resources developed by both professionals and end users in various application fields. The word *grid* is defined as "a system or structure for combining distributed resources; an open standard protocol is generally used to create high quality services." Our approach, applying the grid concept to ensure the collaboration of language services, has not been tried before.

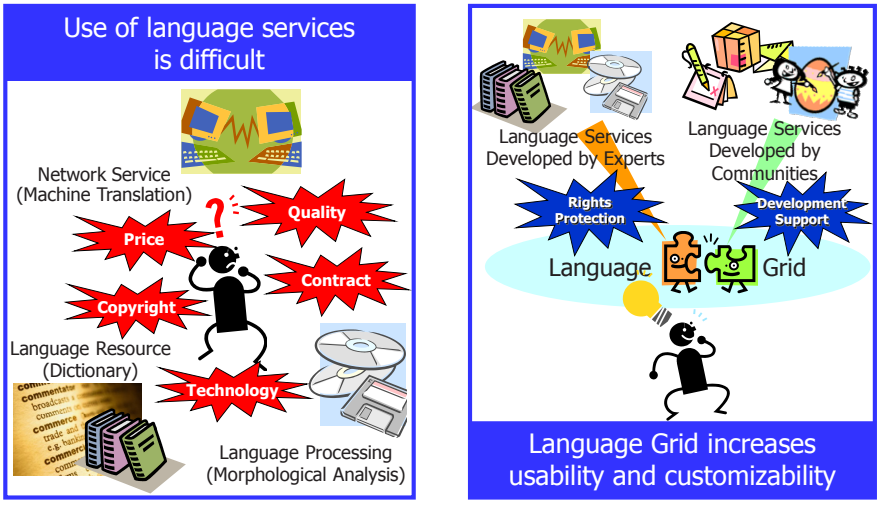

Fig. 1.3 Role of the Language Grid

To realize the Language Grid, *collective intelligence* is a promising solution, since it involves connecting people and computers so that collectively they act more intelligently than any individual or computer (Levy 1999) (Gruber 2008). Recent systems like Wikipedia are successful examples of content-based collective intelligence on the Web. For the Language Grid, however, we propose the *service-oriented collective intelligence* approach. Although the Language Grid lies in the domain of language services, it actually reveals some general problems in open service environments: how to collect and share services, and produce new

services on the Internet. In content-based collective intelligence, contents are always provided by either discarding the intellectual rights or accepting common licenses for the contents. However, the service-oriented approach should handle intellectual property rights issues so as to support service providers to protect their contents and provide services based on their own policies.

Fig. 1.4 illustrates the design concept of the Language Grid. The platform allows users to register services and share them. Major stakeholders of the Language Grid fall into three categories: *service grid operator*, *service provider* and *service user*. Service grid operator manages the Language Grid and controls language resources and services. Service provider provides language services such as machine translations, part-of-speech taggers, dependency parsers, dictionaries, and parallel texts and registers them in the Language Grid. Service user invokes registered language services for their intercultural activities. Note that stakeholders are not individuals but groups like research units in universities, and that a single group can act as two different stakeholders: service provider and service user.

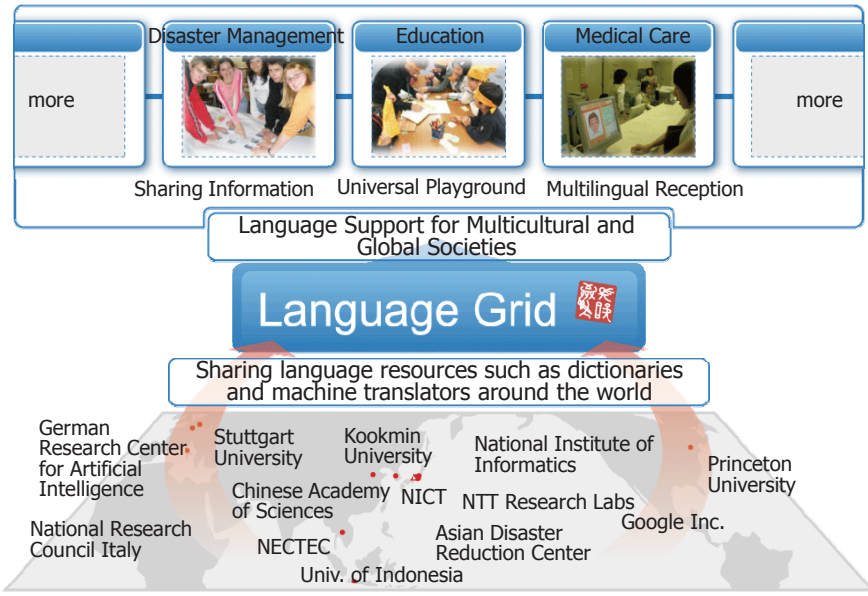

Fig. 1.4 Design concept of the Language Grid

Conceptually, the Language Grid has two main structures: a *horizontal grid* and a *vertical grid*. The horizontal grid concerns the combination of existing bilingual dictionaries or machine translation systems for various languages. The vertical grid concerns specific scenarios of intercultural collaboration activities, which require customized language services including jargon handling.

Among published studies, EuroWordNet (Vossen 1998) and Global WordNet Grid (Fellbaum and Vossen 2007) are pioneers in using word semantics to connect dictionaries in different languages. However, the Language Grid is an attempt to

build a platform that can combine language services provided by stakeholders with different incentives. Therefore, standardization of language services becomes quite important (Calzolari et al. 2002). There also exist several efforts to pipeline language processing programs: *Heart of Gold* (Callmeier et al. 2004) and *UIMA* (Ferrucci and Lally 2004). They aim at pipelining various language processing programs efficiently, but the Language Grid is more application-oriented and focuses on managing the intellectual property rights associated with language resources. Since the motivations are orthogonal, we have bridged Heart of Gold and the Language Grid (Bramantoro et al. 2008), and will apply the results to UIMA.

1.3.2 Service Layers of the Language Grid

As shown in Fig. 1.5, the Language Grid consists of the following four service layers. The bottom layer, called *P2P Grid Layer,* aims at connecting two kinds of servers (*core nodes* and *service nodes*). Core nodes manage all requests to language services, while service nodes actually invoke the atomic services. If the requested service is a composite one, core nodes invoke the corresponding Web service workflow that includes one or more atomic services. Registered information of language services is shared among all core nodes. The same services are provided, regardless of which core node receives the request. The core nodes also control access to services to fulfill the usage conditions set by the service providers. Service providers can access the usage statistics of the services they provide using a system called the *Language Grid Service Manager*.

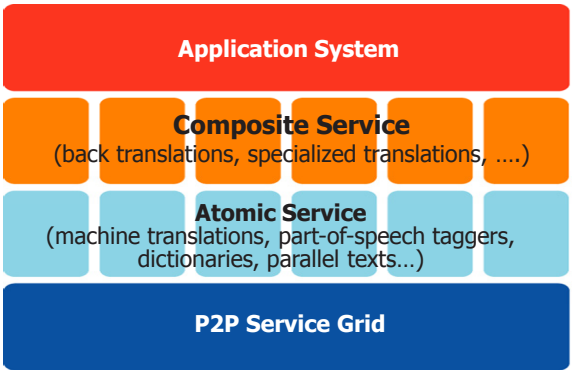

Fig. 1.5 Service layers

The second layer is called the *Atomic Service Layer*. In this layer, any user can add new language resources to the Language Grid. A Web service that corresponds to a language resource is called an *atomic service*. Each language resource is *wrapped* to develop an atomic service. The third layer is the *Composite Service Layer*. Atomic language services can be composed by Web service workflows. A

service described by a workflow is called a *composite service*. Various composite services have been made available up to now, including back-translations and specialized translations. For example, specialized translation can be realized using several atomic services, such as machine translators, part-of-speech taggers, and domain-specific dictionaries. BPEL4WS and Java-based scenarios are used to describe workflows. Currently, more than 90 atomic and composite language services are being shared via the Language Grid with standard interfaces. Table 1.1 lists all types of language services currently available in the Language Grid.

Table 1.1 Language services provided by the Language Grid

Service Category	Service Type	Number of Services
Translation	Translation Service	21
	Domain-Specific Translation Service	2
	Multilingual Mixed Document Translation Service*	0
	Back Translation Service	1
	Multi-hop Translation Service	2
Paraphrase	Paraphrasing Service*	0
	Transliteration Service*	0
Dictionary	Multilingual Dictionary Service	7
	Multilingual Dictionary Service with Longest Match	22
	Concept Dictionary Service	4
	Pictogram Dictionary Service	1
	Multimedia Dictionary Service*	0
	Multilingual Glossary Service*	0
	Dictionary Creation Support Service*	0
Corpus	Parallel Corpus Service	20
	Dialog Parallel Corpus Service	1
	Template Parallel Corpus Service*	0
Analysis	Morphological Analysis Service	7
	Dependency Parsing Service	2
	Similarity Calculation Service*	0
	Language Identification Service*	0
Speech	Text To Speech Service	1
	Speech Recognition Service*	0
Other	Structural Alignment Creation Service*	0
Meta Service	Service Management Service	1

Service types marked with * are currently under development.

To realize the second and third layers of the Language Grid, Web service technologies including *language service ontology*, *horizontal service composition* and *service supervision* have been developed to enable the collaboration needed among language services. Language service ontology is a technology to define standard language service APIs in a hierarchical way so that end users are pro-

vided with simple interfaces while professionals can access more complex interfaces (Hayashi et al. 2008). For horizontal service composition, we apply constraint optimization algorithms to select the appropriate services and thus satisfy QoS requirements (Ben Hassine et al. 2006). To compose machine translators working on the same document or conversation, *context-aware service composition* is proposed: multiple translations are coordinated to determine the meanings of words consistently (Tanaka R et al. 2009). Service supervision, on the other hand, is a runtime technology to monitor and modify the process of composite services (Tanaka M et al. 2009).

Different types of *Application Systems* including collaboration tools have been developed on the top layer. *Language Grid Playground* provides easy access through a Web browser to the Language Grid to try a variety of registered language services. Examples of real-world challenges, such as the creation of community dictionaries, or real-world application of the Language Grid technologies, are also introduced through this website. *Language Grid Toolbox*, on the other hand, is a collection of modules to support multilingual communication in a community. Users can install this software on their servers to offer services, such as multilingual BBS and multilingual dictionary creation. Toolbox is provided as open source software. Therefore, the functions of Toolbox can be extended to meet the requirements of user communities. Furthermore, by using registered language services, existing communication tools can introduce multilingual functions easily. For instance, popular collaboration tools including LiquidThreads and NOTA have been successfully multilingualized.

1.4 User Involvement for Customization

1.4.1 Power of Customization

Computer scientists help to overcome language barriers by creating technologies as language services based on *generalization* of various language phenomena; user communities can then customize and use those technologies to fit their own context by composing language services. There are two reasons why *customization* is a major goal for the Language Grid.

First, machine translators are half-products. The obvious customization step is to combine multilingual dictionaries with machine translators. The provider of those dictionaries does not have to be a research institute or a university. Organizations that are conducting intercultural activities can also register their own multilingual dictionaries. The major difference between machine translation on the Language Grid and a conventional translation system on the Internet is that the users themselves can improve the quality of translation. For example, users can use the registered parallel texts in the translation process. When a user enters a sentence, examples with meanings similar to the entered sentence will appear automatically. If the user is unable to find the intended expression, machine translation is then executed. In this case, a dictionary registered by the user also helps to im-

prove the quality of translation. If the quality of translation is not good enough, however, another user in the multilingual community might manually correct the translation results. The corrected parallel texts are accumulated so that the machine translator can learn from them. This becomes possible when the multilingual community members share their context. In this way, machine-translation-mediated communication might work better in high-context multicultural communities, such as an NPO/NGO working on particular international issues.

Fig. 1.6 Shared screen multilingual chat system for junior high school

Second, we often observe that the success of a multilingual tool in one situation does not guarantee its success in another. Let us examine an example of customized environment for intercultural collaboration. Japan now has an increasing number of students who are non-native Japanese speakers, and most teachers have

a problem in communicating with the foreign students and their parents. We provided a multilingual chat system for a distance meeting as a quick solution, but the system does not work well in face-to-face meetings. Therefore, we developed the service in which users can chat on the same screen. The support site, called a *shared screen multilingual chat system* (see Fig. 1.6), was designed specifically for this situation; students, parents and teachers can chat while looking at the same display. They can input text in their mother tongue, translate the sentence, check the back translation, and post it to the log area at the top of the page. In addition, users can register terms used in the school into the user dictionary, which makes the translation result more correct. This service also provides auto-completion using the parallel texts provided by the city office in charge of the school.

Though extensive dictionaries will help us to find the correct translations of given words, we seldom see people use dictionaries in a conversation. Since language is used everywhere in our daily life, we need customized tools for various situations. You may think that it is not possible to develop such tools customized for different user communities, and may claim that this is the reason that computer scientists provide a generalized solution to cover many different situations. However, the approach taken by the Language Grid is totally opposite. We try to create an environment that allows users to easily develop their own multilingual environments: *the language barrier created by billions of people can be overcome by those billions of people*. For example, the site in Fig. 1.6 was developed in two weeks by three master students. The example shows how quickly a customized multilingual environment can be created by using the language services provided via the Language Grid.

1.4.2 Participatory Design Project

To realize user involvement for customization, we organized a participatory design project that stressed collaboration among researchers, operators and users. At the beginning of the project, in parallel with forming the research project, we established the *Language Grid Association*, a user group of the Language Grid, to conduct multilingual activities on intercultural collaboration. The association is a loosely coupled organization formed by collaboration among industry, government, academia, and citizens with the goal of guiding the development of the Language Grid. Sixteen organizations including laboratories of universities, research institutes, and NPO/NGOs participated in the association. After development of the server software, operation of the Language Grid was commenced. Fig. 1.7 shows the three related organizations. NICT is working on R&D and provides software to the Operation Center run by Kyoto University. Language Grid Association uses the resulting language services and provides feedback to R&D.

The association consists of various SIGs (Special Interest Groups) such as research groups or projects, each of which aims to accumulate use cases and best practices. SIGs can be classified as *creating language services*, *creating collaboration tools* and *supporting multicultural activities* (Sakai et al 2008). Each SIG

creates and shares technologies for using language services registered with the Language Grid. NPO/NGOs, schools and other nonprofit sectors have started to play a central role in breaking down the language barriers. Their activities cover a broad range of fields, including disaster management, education, and medical care (Ishida 2010).

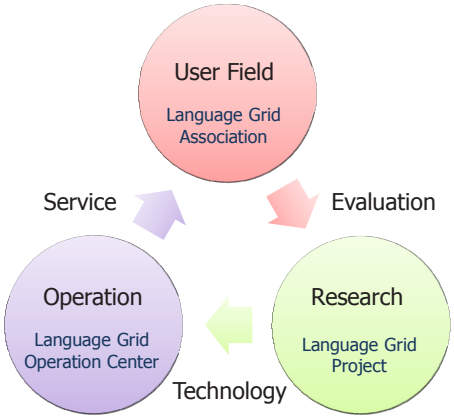

Fig. 1.7 Participatory design project

1.5 Conclusion

The Language Grid is an infrastructure that allows end-users to create new language services for their intercultural collaboration activities. This chapter explained how the Language Grid increases the accessibility and usability of online language resources. Using the Language Grid, various kinds of intercultural activities have begun at hospital receptions, local schools, shopping streets and so on (Ishida et al. 2007) (Fussell et al. 2009).

In general, this chapter proposed the approach of using service-oriented collective intelligence to support the collection, sharing, and production of new services on the Internet while dealing with the issues of intellectual property rights. The main contributions of the proposed approach include the following two aspects.

- *Service architecture*: We developed the service architecture for the Language Grid, including layers of P2P grid infrastructure, atomic services, composite services, and application systems. The proposed architecture applies the service-oriented collective intelligence approach, where language resources including data and software are wrapped as Web services so that users can easily share and combine language resources for creating their own multilingual environment.
- *User involvement*: We proposed a participatory design approach for the formation of service-oriented collective intelligence by bringing the researchers, operators and end users together. Customization of the application systems based

on the Language Grid enables various users and communities to participate in the creation of new multilingual environments.

Our proposed service architecture can be applied to far more than just the language domain. In other domains, the proposed service architecture has been used or is being considered for use in the pervasive computing domain, e-learning domain and so on. In the pervasive computing domain, the open smart classroom has been developed (Suo et al. 2009). In the e-learning domain, open courseware services developed by different organizations can be provided and shared. To provide free usage in various domains, the Language Grid has been released as open source software.

Acknowledgments The project was carried out based on the collaboration between many people in various organizations. We acknowledge the considerable support of the National Institute of Information and Communications Technology, and the Department of Social Informatics, Kyoto University. We are especially grateful to Wolfgang Wahlster, Nicoletta Calzolari Zamorani, Nancy Ide, Christiane D. Fellbaum, Makoto Nagao, Yuichi Matsushima, Takashi Matsuyama, Hiromitu Wakana, Satoshi Nakamura and Hitoshi Isahara for giving advice and encouragement to us. This work was supported by Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society, and a Grant-in-Aid for Scientific Research (A) (21240014, 2009-2011) from Japan Society for the Promotion of Science.

References

Ben Hassine A, Matsubara S, Ishida T (2006) A constraint-based approach to horizontal web service composition. International Semantic Web Conference (ISWC-06), Lecture Notes in Computer Science 4273, Springer: 130-143

Bramantoro A, Tanaka M, Murakami Y, Schäfer U, Ishida T (2008) A hybrid integrated architecture for language service composition. 2008 IEEE International Conference on Web Services: 345-352

Callmeier U, Eisele A, Schafer U, Siegel M (2004) The deep thought core architecture framework. LREC2004: 1205-1208

Calzolari N, Zampolli A, Lenci A (2002) Towards a standard for a multilingual lexical entry: the EAGLES/ISLE initiative. CICLing2002, Lecture Notes in Computer Science 2276, Springer: 264-279

Fellbaum C, Vossen P (2007) Connecting the universal to the specific: towards the global grid. International Workshop on Intercultural Collaboration, Lecture Notes in Computer Science 4568, Springer: 1-16

Ferrucci D, Lally A (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 10: 327-348

Fussell S, Hinds P, Ishida T, Eds. (2009) Proceedings of the Second International Workshop on Intercultural Collaboration

Gruber T (2008) Collective knowledge systems: where the social web meets the semantic web. Journal of Web Semantics 6(1): 4-13

Hayashi Y, Declerck T, Buitelaar P, Monachini M (2008) Ontologies for a global language infrastructure. International Conference on Global Interoperability for Language Resources: 105-112

Ishida T (2006) Language Grid: an infrastructure for intercultural collaboration. 2006 IEEE/IPSJ Symposium on Applications and the Internet: 96-100

Ishida T, Fussell SR, Vossen PTJM, Eds. (2007) Proceedings of the First International Workshop on Intercultural Collaboration, Lecture Notes in Computer Science 4568, Springer

Ishida T (2010) Intercultural collaboration using machine translation. IEEE Internet Computing 14(1): 30-32

Levy P (1999) Collective intelligence: mankind's emerging world in cyberspace. Basic Books

Nomura S, Ishida T, Yamashita N, Yasuoka M, Funakoshi K (2003) Open source software development with your mother language: intercultural collaboration experiment 2002. 2003 International Conference on Human-Computer Interaction: 1163-1167

Paolillo J, Pimienta D, Prado D (2003) Measuring linguistic diversity on the Internet. UNESCO Institute for Statistics, Montreal, Canada

Sakai S, Gotou M, Tanaka M, Inaba R, Murakami Y, Yoshino T, Hayashi Y, Kitamura Y, Mori Y, Takasaki T, Naya Y, Shigeno A, Matsubara S, Ishida T (2008) Language Grid Association: action research on supporting the multicultural society. 2008 International Conference on Informatics Education and Research for Knowledge-Circulating Society: 55-60

Suo Y, Miyata N, Morikawa H, Ishida T, Shi Y (2009) Open smart classroom: extensible and scalable learning system in smart space using web service technology. IEEE Transactions on Knowledge and Data Engineering 21(6): 814-828

Tanaka M, Ishida T, Murakami Y, Morimoto S (2009) Service supervision: coordinating web services in open environment. 2009 IEEE International Conference on Web Services: 238-245

Tanaka R, Murakami Y, Ishida T (2009) Context-based approach for pivot translation services. International Joint Conference on Artificial Intelligence (IJCAI-09): 1555-1561

Vossen P (1998) Introduction to eurowordnet. Computers and the Humanities 32(2-3): 73-89

Weiss A (2005) The power of collective intelligence. netWorker 9(3): 16-23

Chapter 2
Service Grid Architecture

Yohei Murakami[1], Donghui Lin[1], Masahiro Tanaka[1], Takao Nakaguchi[2], and Toru Ishida[3]

1 National Institute of Information and Communications Technology (NICT), Language Grid Project, 3-5 Hikaridai, Seika-Cho, Soraku-Gun, Kyoto, 619-0289, Japan, e-mail: {yohei, lindh, mtnk}@nict.go.jp

2 NTT Advanced Technology Corporation, 12-1 Ekimae-Honmachi, Kawasaki-Ku, Kanagawa, 210-0007, Japan, e-mail: takao.nakaguchi@ntt-at.co.jp

3 Department of Social Informatics, Kyoto University, Yoshida Honmachi, Sakyoku, Kyoto 606-8501 Japan, e-mail: ishida@i.kyoto-u.ac.jp

Abstract The Language Grid is an infrastructure for enabling users to share language services developed by language specialists and end user communities. Users can also create new services to support their intercultural/multilingual activities by composing language services from a range of providers. Since the Language Grid takes the service-oriented collective intelligence approach, the platform requires the services management to satisfy stakeholders' needs: access control for service providers, dynamic service composition for service users, and service grid composition and system configurability for service grid operators. To realize the Language Grid, this chapter describes the design concept and the system architecture of the platform based on the service grid.

2.1 Introduction

Although there are many language resources (both data and programs) on the Internet (Choukri 2004), most intercultural collaboration activities still lack multilingual support. To overcome language barriers, we aim to construct a novel language infrastructure to improve accessibility and usability of language resources on the Internet. To this end, the Language Grid has been proposed (Ishida 2006). The Language Grid takes a service-oriented collective intelligence approach to sharing language resources and creating new services to support their intercultural/multilingual activities by combining language resources.

In previous works, many efforts have been made to combine language resources, such as UIMA (Ferrucci and Lally 2004), GATE (Cunningham et al. 2002), D-Spin (Boehlke 2009), Hart of Gold (Callmeier et al. 2004), and CLARIN (Varadi et al. 2008). Their purpose is to analyze a large amount of text data by linguistic processing pipelines. These pipelines consist of language resources, most

T. Ishida (ed.), *The Language Grid: Service-Oriented Collective Intelligence for Language Resource Interoperability*, Cognitive Technologies, DOI 10.1007/978-3-642-21178-2_2,

of which are provided as open sources by universities and research institutes. Users can thus collect language resources and freely combine them on those frameworks without considering other stakeholders.

Different from the above frameworks, the purpose of the Language Grid is to multilingualize texts for supporting intercultural collaboration by service workflows. A workflow combines language resources associated with complex intellectual property issues, such as machine translators, parallel corpora, and bilingual dictionaries. These resources are provided by service providers who want to protect their ownership, and used by service users who need a part of the resources. Therefore, the Language Grid must coordinate these stakeholders' motivations. That is, it requires language service management to satisfy the following stakeholders' needs as well as language service composition for service users.

- Protecting intellectual properties of resources: Some service providers can agree on providing their services if they can retain ownership of their resources and specify the extent that service users utilize the services. For detecting fraudulent usage, they also want to know what their service is used for.
- Utilizing necessary services when needed: Service users want to utilize necessary services when needed, but not own the resources. Moreover, they may want to customize composite services for their goals by freely combining services.
- Configuring platform according to operation models: Operators create various operation models to meet stakeholders' needs. To fit their platforms to their operation models, they need to optimize system configuration. In addition, by connecting their platforms, they want to allow service users to share and invoke services on other platforms.

The above requirements are not only inherent in the Language Grid and language resources, but in any system composing services provided by others. Here we call the platform to share and compose services provided by different providers as the *service grid*, and design the service grid architecture. The service grid is a general platform independent of specific service domains so that it can be applied to a specific domain by defining services specific to the domain. For example, the Language Grid is a service grid specific to the language resource domain. Firstly, based on these requirements, this chapter clarifies functions that the service grid should provide, and explains its design concept and system architecture. Furthermore, we validate the service grid architecture by using it as basis for constructing the Language Grid.

2.2 Design Concept

The purpose of the service grid is to accumulate services and compose them. To realize the service grid, system architecture should be designed to satisfy different requirements from the stakeholders. Therefore, this section summarizes require-

ments of each of the stakeholders, and clarifies the required functions of the service grid.

2.2.1 Requirements

Service providers demand prevention of data leaks and fraudulent usage of resources because the resources represent intellectual properties. Specifically, the service providers want to deploy their services on their servers and provide their services following their provision policies, but not publish their resources under a common license, like Wikipedia. Furthermore, to check whether service users employ their services properly, they may want to know when their services are accessed and who accesses them.

On the other hand, service users prefer flexibility in customizing services and convenience in invoking the services to acquiring ownership of the resources. This is because they want to concentrate on developing application systems by reducing the resource maintenance cost. Specifically, they need to access the services through standard Web service technologies over HTTP. Moreover, they also need to create composite services freely and change service combinations.

Finally, service grid operators require flexibility of system configuration so that they can adapt the configuration to stakeholders' incentives. For example, the operators operate the service grid on a single cluster of machines by collecting services if the provision policies of the services are relaxed. Meanwhile, they operate the service grid in a distributed environment by deploying services on each provider's server if the provision policies of the services are too strict. In the former case, the operators place high priority on performance of services. In the latter case, they put priority on resource security. Further, they may want to expand available services by allowing their users to access services on other service grids.

2.2.2 Functions

The service grid platform should provide the following functions extracted from the stakeholders' requirements in the previous subsection.

(1) Service access control and monitoring: Service providers can set out their provision policies defining the terms of service use, and the platform controls access to the services according to these policies. For instance, restrictions on users who may be licensed to use the service, on the purpose for which the service may be used, and on the number of times the service may be accessed, the amount of data that may be transferred from the service, and so on. Furthermore, the platform accumulates service request messages and service response messages as access logs, and enables service providers to monitor service invocation histories and the status of