

Language Service Infrastructure on the Web: The Language Grid

Toru Ishida, Kyoto University

Yohei Murakami, Ritsumeikan University

Donghui Lin, Kyoto University

Takao Nakaguchi, The Kyoto College of Graduate Studies for Informatics

Masayuki Otani, Kindai University

Globalization increasingly demands multilingual communication on the Internet, as well as in local communities. To create customized collaboration tools to support multilingual communities, the authors' Language Grid, established ten years ago, has been improving Web-based services to communities throughout the world by providing highly adaptable infrastructure and access to a wide variety of language resources.

Globalization has increased multilingual communication in the virtual space of the Internet. It has also triggered large-scale human migration, and thus created a huge demand for multilingual communication in local communities. Immediately after 9/11, we conducted the intercultural

collaboration experiment (ICE) among Chinese, Japanese, Korean, and Malay universities in an attempt to use machine translation to overcome language barriers.¹ In this trial, open source software was jointly developed through the Internet. Since the experiment pursued collaboration across language borders, participants were

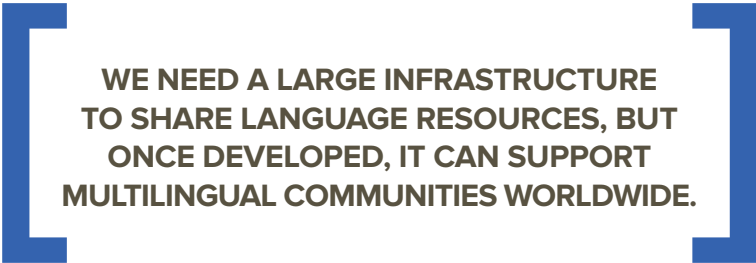
requested to communicate only in their native tongue with machine translation. To the best of our knowledge, this was the first attempt to use machine translation in a large-scale collaboration task. The experiment spanned six months, including software design to submit a system proposal for implementation and development of the software to be released.

We then analyzed the more than 15,000 translated sentences collected during the experiment. Recent progress in machine translation is tremendous, but because machines cannot learn the various regional jargon used in specific domains, which often play a key role in communication, their mistranslation disrupts the entire conversation. Furthermore, while the mental models of the translation machines were hard to obtain, we did find that the participants had two repair patterns: in *self-initiated repair*, the user adapted to machine translation by repeatedly altering the input text to improve the translation quality; and in *other-initiated repair*, collaborative translation was conducted to clarify the intentions of the remarks.

We learned how important it is to customize machine translators. Since users cannot modify the translation software, there is no way for them to combine machine translators and dictionaries specific to their community. It is also difficult to collect and integrate language tools: machine translators and dictionaries. It is not easy for users to evaluate translations, investigate contracts, and to select companies. We also found that the participants continued their attempts to improve the quality of translation. This indicates the value of a flexible multilingual environment that can continuously reflect users' demands.

In summary, even for small multilingual communities, regardless whether virtual or local, it is essential to create a sociotechnical system² that can share language resources regardless of source. We need a fairly large infrastructure to share language resources, but once developed, it can support multilingual communities worldwide. Since a wide variety of autonomous stakeholders will provide an array of technical resources, the infrastructure needs to be able to resolve the conflict between resource protection of service providers and the usability concerns of service consumers, to hide technical

catalogue. More than one thousand resources are downloadable, including lexicons, corpora, multimodal resources, and the like. In the US, LDC (Linguistic Data Consortium) was formed in 1992. The LDC Catalog contains 841 resources, such as corpora, images, spoken resources, and tools to support annotation tasks. There are also a variety of public and commercial language resources already available online. Unfortunately, the variety of contracts, intellectual property rights and application interfaces often become barriers to the usage of those resources.



**WE NEED A LARGE INFRASTRUCTURE
TO SHARE LANGUAGE RESOURCES, BUT
ONCE DEVELOPED, IT CAN SUPPORT
MULTILINGUAL COMMUNITIES WORLDWIDE.**

differences by introducing standardized service interfaces, and to lower the operation and maintenance costs for sharing language resources. This article describes a sociotechnical approach to develop such an infrastructure, called the Language Grid.³

LANGUAGE RESOURCES

Two organizations have been established in Europe and the US with responsibility to share language resources. In Europe, ELRA (European Language Resources Association) was founded in 1995. It offers the ELRA Catalogue, a repository of language resources and their quality assessments. Users can select their required resources by reference to the

Language resources can be divided into two categories: software and data. Concerning software, machine translators are often commercial and sold as packages or services on the Web. Though several global portals seem to provide free translation services, when accessed via their APIs, charges are levied or frequency of use is restricted. On the other hand, natural language analysis software—such as part-of-speech taggers, parsers, and so on—has been developed by university laboratories and research institutes. Some of those programs are published under an open source license, provided for free without any restriction, while others are provided only for research or nonprofit use.

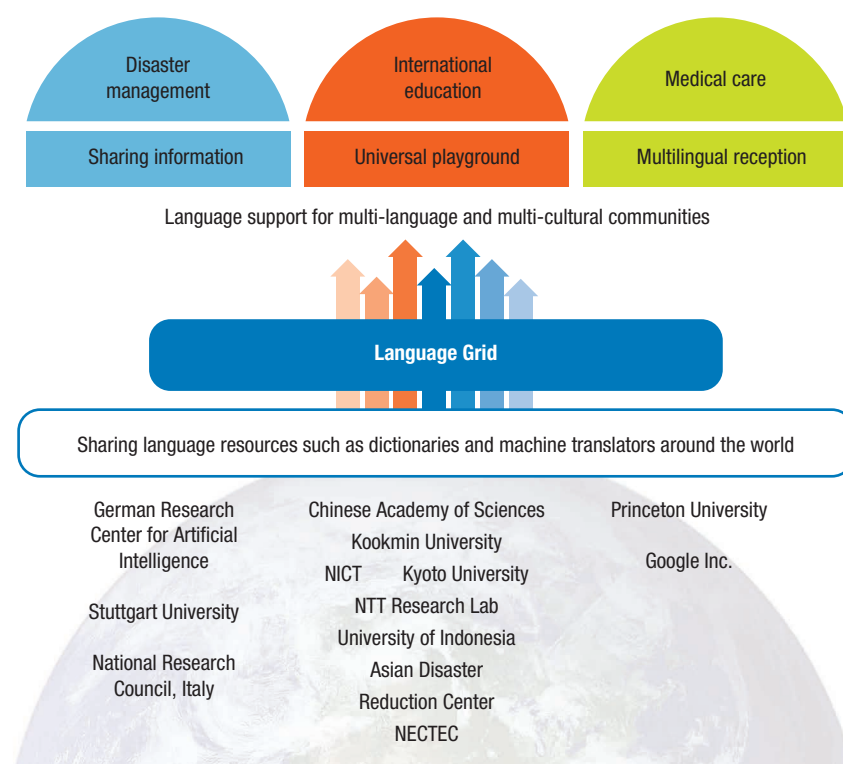


FIGURE 1. Concept of the Language Grid. Various applications can use the language services provided by research institutions and companies worldwide.

The usage conditions of data are even more varied. Dictionaries require huge efforts to create, but their intellectual property rights are not straightforward. The creativity of a dictionary usually determines the strength of its copyright. If the dictionary cannot be protected by copyright, and thus must secure compensation, the publisher is hesitant to provide data for free even for academic use.

The same problems exist for corpora. Copyright issues on data extracted from documents cause complex problems. Statistical data can be used freely but distributed representations in neural networks, which may create sentences very close to the original ones, are problematic. As different

authors will prepare different licenses for their resources, users wanting to access various resources are forced to understand a variety of licenses. In addition, universities and research institutes seldom provide concrete documents on their copyright policies, but this does not mean they offer resources unconditionally. Users often have to negotiate with the authors to understand their usage conditions.

The size of language resources affects the quality of services such as machine translations. One trend in the era of big data is to pay scant attention to small data. As far as language resources are concerned, high-resource languages such as English and Chinese attract more commercial attention,

whereas low-resource languages are discussed more intensively in research communities. This is because languages matter not only for businesses, but also for cultures.

LANGUAGE RESOURCE SHARING AS SERVICE

In 2006, we started a project called the Language Grid, see Figure 1. Though there was no clear concept of language services in academia, we started operating an experimental infrastructure in 2007 to accumulate and share language resources as Web services.⁴ Note that our intention was not to gather language resources like ELRA or LDC, but rather to request that various institutions provide their language resources as services, and we would connect them to our servers.

The idea of the service grid architecture is based on the concepts of *fragmentation*, which provides various language services, and *recombination*, which realizes a customizable language environment. The slogan of our project is “from language resources to language services.” The implementation of this idea created a full range of customized language environments for different types of user activities.

In *fragmentation*, language resources are wrapped with standardized interfaces. We wrap existing language resources as atomic language services and enable users to share them. For example, the interfaces of existing machine translators are not identical; some machine translators require only a sentence to be translated and others require a language code as well. We wrap every translator with a common interface in order to allow users to use machine translators without considering the interfaces involved. Human interpreters can be

also wrapped with the standard translation interface, so that users do not have to distinguish between machines and humans.

In *recombination*, atomic language services are combined to create new services. Users can switch translation services flexibly to ensure the quality of services: machine translators can be used for chat text, which requires quick translation, while human interpreters are preferred for documents that demand high translation quality.

Various services computing technologies have been developed over the last decade. For the Language Grid, we invented *horizontal service composition*,⁵ which can select the best atomic service from a set of atomic services to instantiate a given service workflow. Constraint optimization algorithms are applied to maximize the quality of service (QoS) of the workflow. This technology is most suited to language services, as many alternative services exist with a common interface. *Policy-aware optimization of parallel execution*⁶ was also invented to predict optimal degrees of parallelism for atomic and composite services. This technology is extremely useful in the current situation in which commercial portals provide different machine translation services with different policies.

STAKEHOLDERS

We proposed the service grid architecture to increase the usability of language resources, and to decrease the risk to providers in opening their resources. By wrapping resources as services, providers can control their intellectual property rights.

To share language resources as services, we designed the service grid architecture shown in Figure 2. However, it is essential to define

stakeholders, their roles and the social protocol among them.⁷ We call every stakeholder related to service grids a *service grid user*; each can take one or more roles in the following three categories.

- › *Service provider*: wraps language resources into language services, and deploys them on a grid. When registering services, access control policies can be specified for each service.
- › *Service consumer*: invokes registered language services from an application system. When

view, protection of intellectual property rights is critical. To satisfy such demands, we classify service usage into the following three categories, so that providers can restrict service usage: *nonprofit use*, which allows use of services in public, nonprofit, or private settings; *research use*, which allows use of services when they are intended to advance the field, not for commercial profit; and *for-profit use*, which allows use of services for commercial profit regardless of the type of organization.

The above classifications can also be applied to individual users, except

WE PROPOSED THE SERVICE GRID ARCHITECTURE TO INCREASE THE USABILITY OF LANGUAGE RESOURCES.

invoking a composite language service, the request is sent to a workflow engine, which executes a workflow that combines one or more atomic language services.

- › *Grid operator*: manages and controls a service grid for service providers and consumers. Each service grid has a grid operator.

Most service providers are universities or companies, whereas service consumers are often nonprofit organizations or the public.

The institutional agreement we designed reflects the intentions of the three roles of service grid users. From the service providers' point of

private use, which is always classified as nonprofit. Note that even in for-profit organizations, socially responsible activities are classified as nonprofit use. This is because such activities are often conducted with public institutions or nonprofit organizations. Conversely, the activities of public institutions or nonprofit organizations for commercial profit are classified as for-profit use.

From the service consumers' point of view, the important issue is whether or not they can develop an application system by combining language services. There are two types of application systems: one provides services to anonymous people via the Web, and the other provides services via specific terminals.

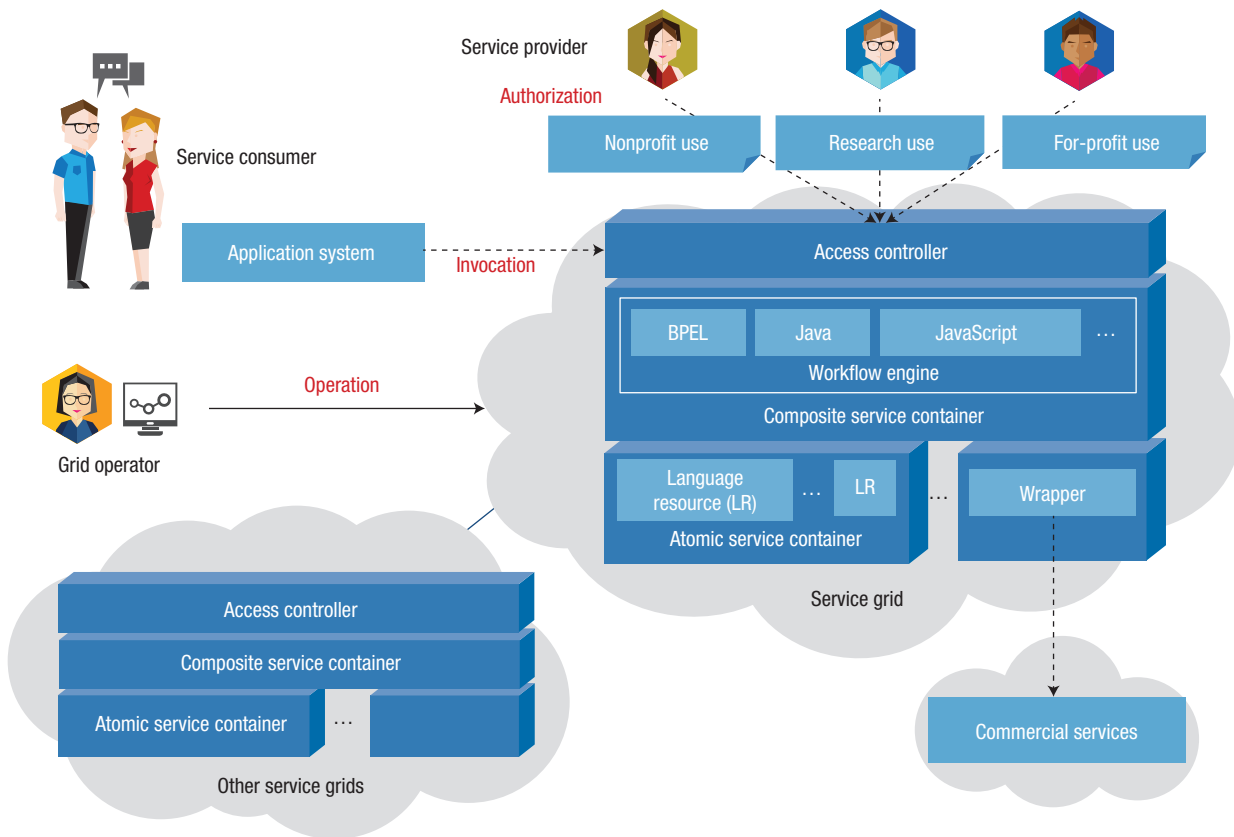


FIGURE 2. Service grid architecture. Each service grid user takes one or more roles: service provider, service consumer, and grid operator.

The former includes the cases in which local governments create a QA site for foreign citizens. The latter includes the cases at reception desks of hospitals to support foreign patients. In both cases, the critical issue for service consumers is how the application system can control service usage.

When service providers register services, copyrights and other intellectual property rights should be specified clearly by using the above three categories. The service consumer self-identifies as one of the categories when using a service, so that the access controller can decide whether or not

the consumer can use the service. Of course, service providers can specify their own terms, regardless of which of the above categories is specified. This option increases the satisfaction of service providers, while decreasing the service usability.

FEDERATED OPERATION

Our ten-year operation of the Language Grid has yielded a critical insight: the top three languages supported by the Kyoto grid operator, namely, English, Chinese, and Japanese, occupy more than half of the language services. This reflects the

difficulty of the grid operator in Kyoto in reaching service providers in other countries. The locality, caused by geographical conditions, became a driving force for creating a new service grid. To encourage local development of language resources, we tackled this bias: grid operators must be globally dispersed, operate local grids, and exchange services with one another. The collaboration of grid operators can be realized by *federated operation*.⁸

We classify federated operation into two types. In *centralized federation*, the grid operators form an association to create and maintain the affiliations

among them. This yields flexibility in deciding or altering the affiliation regulations, but maintaining the federal association costs a lot. In *distributed federation*, on the other hand, a service grid user is allowed to create a new service grid and to become its operator. Since grid operators are connected in a peer-to-peer fashion, the maintenance costs of the entire network can be small. The Language Grid is an experimental system, and thus follows the distributed federation approach; university laboratories can create a large network.

The implementation of a distributed federation is shown in Figure 3. An *affiliated operator*, a service grid user of the original grid operator, is the grid operator of its own service grid. An *affiliated user* is a service grid user of an affiliated operator. The main idea of distributed federation is to allow affiliated users to use services registered to the original service grid. If the original grid operator is also a service grid user of the affiliated operator, the two operators establish an equal partnership, namely, a bidirectional affiliation. When both operators focus on language services, they may prefer bidirectional affiliation. However, the service grid can be applied to service domains other than languages: one operator focuses on language services while the other focuses on tourism, the latter may use the former's services, but not vice versa.

Since our grid server software is published under open source licenses, any organization can start its own service grid. From the service providers' point of view, however, they may want to avoid the case that the range of service consumers expands automatically through the incremental affiliation of grid operators. Thus, it is necessary for

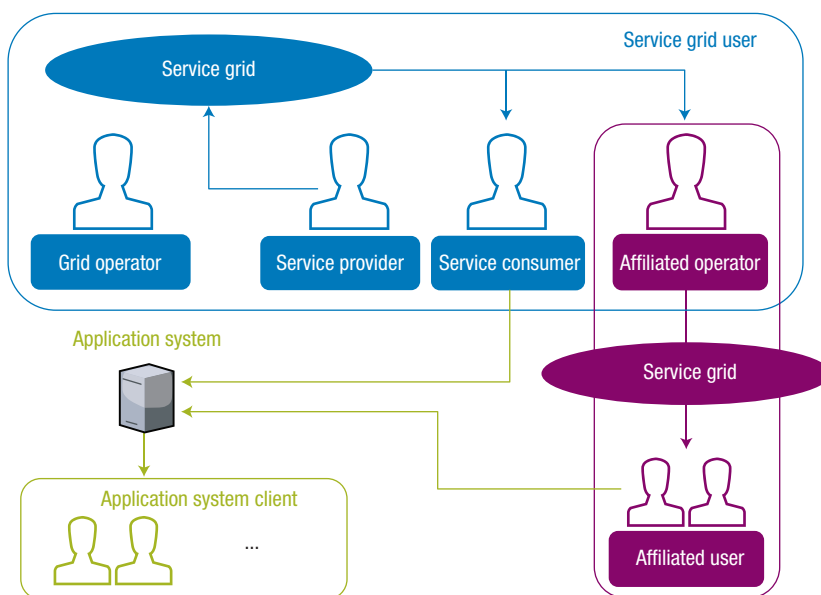


FIGURE 3. Federated operation. Service grid users are licensed by a grid operator to access the service grid under an institutional agreement.

service providers to specify whether or not they allow affiliated users to use their services.

We developed an institutional agreement that includes the above details for realizing distributed federations. Once service grid users conclude the agreement with a grid operator, no additional contract is needed to become an affiliated operator. In this way, we suppress additional overheads while extending the network of service grids. Under the guise of distributed federation, we have created a network of operation centers to cover various Asian languages. NECTEC, University of Indonesia, and Xingjiang University became affiliated grid operators of the Kyoto grid operator. Operation centers were opened in Bangkok in 2010, Jakarta in 2011, and Urumqi in 2014, and have connected themselves to the Kyoto operation

center to share a variety of services in Asian languages.

REAL-WORLD APPLICATION

The Language Grid has supported various kinds of intercultural activities, including at hospital reception desks, local schools, shopping districts, and so on. We take a community-based approach⁹ to ensure the quality of services. A typical example is given below.

From 2011 to 2014, we worked on deploying our service grid architecture to create a real-world application in Vietnam. Two major goals were dealt with in this project: low rice productivity and the environmental burdens caused by the excessive use of agrichemicals. We conducted a four-month experiment each year in Vinh Long Province, which is located in Vietnam's Mekong Delta. The goal

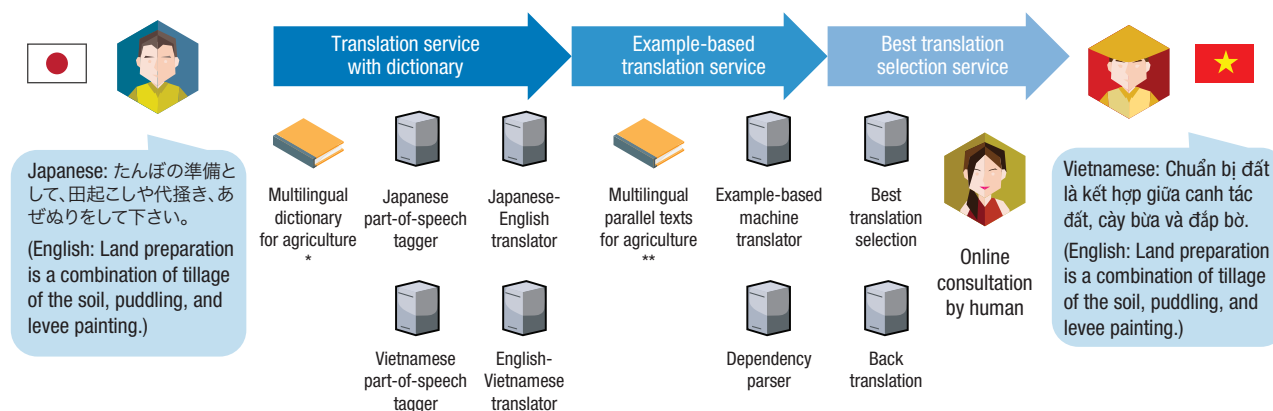


FIGURE 4. Language service composition. Language resources (dictionaries, parallel texts, part-of-speech taggers, machine translators, human interpreters, and so on) were combined to create a communication channel. Google Translate was used for English-Vietnamese translation, while J-Server of Kodensha Inc., was used for Japanese-English translation. Dictionaries and parallel texts for agriculture were created by this project. *Multilingual Dictionary for Agriculture was jointly developed by NPO Pangaea, Japan National Agriculture Research Center, Vietnam MARD. Entry Number: 3,099 (Sep. 2014). **Multilingual Parallel Texts for Agriculture was jointly developed by NPO Pangaea, Japan National Agriculture Research Center, Vietnam MARD. Entry Number: 2,485 (Sep. 2014).

was to provide timely and appropriate agriculture knowledge in rice harvesting for Vietnamese farmers by Japanese experts. Since Japanese experts cannot physically travel to all rural areas, they were highly motivated to use information technology. However, because of low literacy rate in the Vinh Long Province, the farmers had difficulties using computers and indeed in reading/writing messages. With agriculture experts and NGO staff, we thus invented the youth-mediated communication (YMC) model, where children act as mediators and bridge the gaps in language, knowledge, and cultures between experts and farmers.¹⁰

Figure 4 illustrates a sample workflow in which functionalities are enhanced by combining a variety of atomic language services so that Japanese agricultural experts can transfer their knowledge to Vietnamese farmers. The leftmost workflow

cascades Japanese-English and English-Vietnamese translators. To translate jargon into appropriate words, the workflow divides the input sentences into words by using part-of-speech taggers, and finds the jargon in the input sentences by using multilingual dictionaries for agriculture. To improve the translation quality, the middle workflow, we trained example-based machine translators with multilingual parallel texts for agriculture. This yielded two translation services and raised the problem of determining which one would be best. The rightmost workflow comprises back translation such as Japanese-Vietnamese-Japanese translation and compares original and back-translated sentences. It selects the translation service whose back-translated sentences are most similar to the original ones. Finally, if the quality of translation is still not satisfactory, human interpreters are consulted online.

Unfortunately, the available machine translation technologies are insufficient, especially when translators are cascaded. For better translation, we invited human bridgers to join the project. Typically, students of agriculture departments, these bridgers were highly motivated to assist this project and to acquire professional knowledge. Since machine translators are designed to input correct sentences, the bridgers often pre- and post-edited the input sentences in cascaded translation, which in this case were English sentences. Figure 5 shows the workflow of bridgers. We observed two repair patterns here: self-initiated repair was realized by the interaction between a bridger and translation services, whereas other-initiated repair was realized by interaction among bridgers.

Intercultural collaboration in the real world involves more than just dealing with language technologies, it is also a community-building process.

We developed a customizable multilingual tool for this project to flexibly support project formation. The YMC model is a breakthrough idea, but requires collaboration among different stakeholders. At the beginning of the YMC project, we focused on multilingual communication services for knowledge transfer from experts to farmers. Children and bridgers then joined to refine the communication channel. After the project started, we observed that experts become motivated to understand what was happening in remote paddies.

A complementary service was added in which the farmers are service providers and the experts are consumers. Meanwhile, the local government officers found that the YMC model contributed to the formation of communities in low-literacy areas. Staffs from the local government joined to support the farmers' children in interpreting the expert advice. In this way, the community grew to utilize and to make full use of immature technologies. Figure 6 shows a photo of the community-based project including various stakeholders.

Launched in 2007, the Language Grid now has 183 participating groups from 24 countries sharing 226 language services. Four grid operators are connected with each other. There are 87 bilingual dictionaries and parallel corpora, 24 translators, and 16 part-of-speech taggers provided from Kyoto; 14 Asian WordNets operate out of Bangkok; and Jakarta and Urumqi provide language services for Indonesian and Turkish language families. Recently, our server software has been selected as the basis of a federated grid for

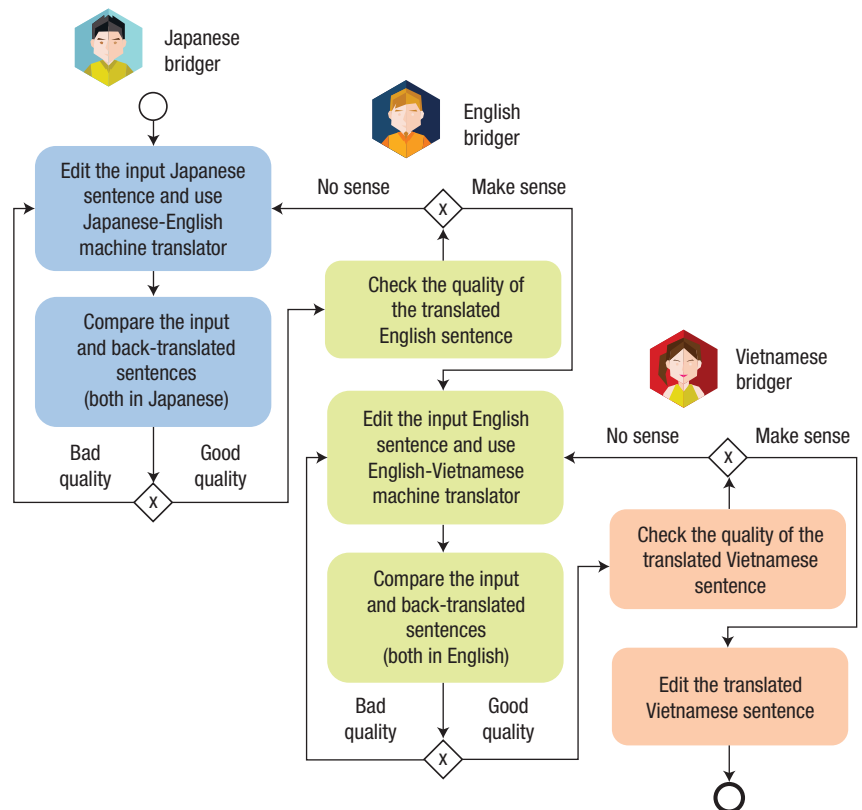


FIGURE 5. Collaborative translation workflow. Bridgers interact with language services and other bridgers to improve the quality of Japanese-Vietnamese translation.

language services, in which Asian, US, and EU projects, supported by ELRA and LDC, will share open source language resources.¹¹

The Language Grid and its field application can be seen as the emergence of a social machine.¹² Lessons learned include not only *fragmentation* and *recombination* of language services, but also the successful design of protocols among humans and machines. We adopt the sociotechnical approach in which stakeholders provide and consume language services. We have confirmed the need for institutional agreements to reduce friction in collaboration, and the need for human involvement, since

machine translation is still imperfect, especially when involving a third language in the middle. Engineers tend to think immature technologies are not yet useful, but if a real need exists, the technologies can encourage stakeholders to create a problem-solving organization, as we have done here through combining language services. As a result, various individuals—from community members to government officials—will play important roles in utilizing language services.

Based on our experience, we intend to pay more attention to low-resource languages. More than 7,000 languages exist around the world. In 2010,



FIGURE 6. Problem solving community with the Language Grid in Vinh Long Province, Vietnam. Various stakeholders including NGO staff, children, bridgers, professors, and government officers (left to right) form a problem-solving organization.

UNESCO released a list of 2,464 endangered languages. To preserve those languages, linguists are actively engaged in describing syntax and vocabularies and recording texts, voices, and movies.¹³ Our contribution is to create services for low-resource languages: automatic generation algorithms have been studied to create dictionaries between minor languages by combining existing dictionaries via a pivot language,¹⁴ specifically, our algorithm can create a Uyghur-Kazakh dictionary from Uyghur-Chinese and Kazakh-Chinese dictionaries. We are also developing dictionaries in Indonesia, where 144 languages are endangered. These efforts will increase the usage of low resource languages, and remove them from the endangered list.

ACKNOWLEDGMENTS

We thank collaborators including NPO Pangaea, and agriculture experts in Universities of Tokyo and Mie. This research is partially supported by Service Science, Solutions and Foundation Integrated Research Program from JST RISTEX, and Grant-in-Aid for Scientific Research 24220002, 17H00759 and 17H04706.

REFERENCES

1. S. Nomura et al., "Open Source Software Development with your Mother Language: Intercultural Collaboration Experiment 2002," Proc. Int'l Conf. Human-Computer Interaction (HCII 03), 2003, pp.1163-1167.
2. M.P. Singh, "Norms as a Basis for Governing Sociotechnical Systems," *ACM Trans. on Intelligent Systems and*

Technology, vol. 5, no. 1, pp. 21:1-21:23.

3. T. Ishida, ed, *The Language Grid: Service-oriented Collective Intelligence for Language Resource Interoperability*, Springer, 2011.
4. T. Ishida, "Language Grid: An Infrastructure for Intercultural Collaboration," *Proc. IEEE/IPSJ Symp. on Applications and the Internet*, 2006, pp. 96-100.
5. A.B. Hassine et al., "A Constraint-based Approach to Horizontal Web Service Composition," *Proc. Int'l Semantic Web Conf (ISWC 06)*, 2006, pp. 130-143.
6. M.X. Trang et al., "Policy-Aware Optimization of Parallel Execution of Composite Services," *Proc. IEEE Services Computing Conf. (SCC 15)*, 2015, pp. 106-113.

7. A.K. Chopra and M.P. Singh, "From Social Machines to Social Protocols: Software Engineering Foundations for Sociotechnical Systems," *Proc. Int'l Conf. World Wide Web (WWW 16)*, 2016, pp. 903-914.
8. Y. Murakami et al., "Service Grid Federation Architecture for Heterogeneous Domains," *Proc. IEEE Services Computing Conf. (SCC 12)*, 2012, pp. 539-546.
9. D. Murray-Rust et al., "A Collaboration Model for Community-based Software Development with Social Machines," *Proc. Int'l Conf. Collaborative Computing (CCC 14)*, 2014, pp. 84-93.
10. Y. Mori et al., "Youth Mediated Communication: Agricultural Technology Transfer to Illiterate Farmers through their Children," *Proc. World Conf. on Computers in Agriculture*, 2012.
11. Y. Murakami, D. Lin, and T. Ishida, eds., *Services Computing for Language Resources*, Springer, 2018.
12. J. Hendler and T. Berners-Lee, "From the Semantic Web to Social Machines: A Research Challenge for AI on the World Wide Web," *Artificial Intelligence*, vol. 174, no. 2, 2010, pp. 156-161.
13. L.A. Grenoble and L.J. Whaley, *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press, 2006.
14. M. Wushouer et al., "A Constraint Approach to Pivot-based Bilingual Dictionary Induction," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP 16)*, 2016, vol. 15, no. 1, Article 4.

myCS

Read your subscriptions
through the myCS
publications portal at

<http://mycs.computer.org>

ABOUT THE AUTHORS

TORU ISHIDA is a professor in the Department of Social Informatics, at the Graduate School of Informatics, Kyoto University. He has been working on multi-agent systems and conducting projects including Digital City Kyoto and the Language Grid. Ishida received a PhD in engineering from Kyoto University. Contact him at ishida@i.kyoto-u.ac.jp.

YOHEI MURAKAMI is an associate professor at the Graduate School of Information Science and Engineering, Ritsumeikan University. He has been working on services computing and leading the research and development of the Language Grid to support intercultural collaboration. Murakami received a PhD in informatics from Kyoto University. Contact him at yohei@fc.ritsumei.ac.jp.

DONGHUI LIN is an associate professor in the Department of Social Informatics, at the Graduate School of Informatics, Kyoto University. He has been working on services computing and conducting the research and development of the Language Grid for intercultural collaboration. Lin received a PhD in informatics from Kyoto University. Contact him at lindh@i.kyoto-u.ac.jp.

TAKAO NAKAGUCHI is an associate professor at The Kyoto College of Graduate Studies for Informatics. He has been working on services computing. He developed the Language Grid server software. Nakaguchi received a PhD in informatics from Kyoto University. Contact him at ta_nakaguchi@kcg.edu.

MASAYUKI OTANI is a lecturer in the Department of Informatics, at the Faculty of Science and Engineering, Kindai University. He has been working on the area of multi-agent systems for the internet of things and intercultural collaboration. Otani received a PhD in engineering from the University of Electro-Communications. Contact him at otani@info.kindai.ac.jp.

IT Professional
Technology Solutions for the Enterprise



www.computer.org/itpro