

Master Thesis

Application of Example-Based Machine Translation to Multilingual Community

Supervisor Professor Toru Ishida

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Satoshi MORIMOTO

February 2, 2010

Application of Example-Based Machine Translation to Multilingual Community

Satoshi MORIMOTO

Abstract

This paper describes supports of multilingual communication which becomes a big problem at various places in recent years. Multilingual communication is, for example, a communication between a doctor and a foreign patient in a hospital or a communication between a teacher and a foreign student or his/her parent in a school. And some support to such multilingual community is needed because smooth communication between people who cannot speak common language in the community is impossible by language barrier problem.

To the problem of language barrier, technology of machine translation is being developed and solving the problem. But features and problems when using machine translation in multilingual community, became clear by an experiment of multilingual communication when only using machine translation. One of these problems is that many machine translators cannot translate special terms or peculiar phrases used in a certain community or a certain domain such as medical fields or fields of education. That is, using only machine translate cannot support smooth multilingual communication.

To solve the above problem, the Language Grid Project, which aim is to break the language barrier, has worked on a variety of supports of multilingual communication on a certain community. For example, Translation with Dictionary, which is constructed by combining an existing machine translator with language resources such as dictionaries of special terms and morphological analyzers, can improve translation quality of sentences including domain-specific terms. That is to say the Language Grid has helped users who needs to be supported while multilingual communication to use machine translation well, and is actually used at various places. And many applications for supporting multilingual community using the Language Grid have been developed.

As the above many supports of multilingual community by using the Language Grid have been completed and all of them gave actual results. But when considering further support of multilingual community from now on, I

think customized machine translation to a multilingual community which requires support is needed. The quality of the translation of sentences including domain-specific terms is improved by Translation with Dictionary, which is used widely in activities using Language Grid up to now, but some problems are still left such as that phrases peculiar to a certain community cannot be translated well by only translation with dictionary. In this paper, I will discuss how to customize and apply machine translation for multilingual community support to solve the problem.

When customizing machine translation in a community, translation rules or a bilingual corpus about sentences used in the community are needed. But in this paper the Example-based Machine Translation will be focus on, because when considering the problem of the cost it is known that it is suitable to use the Example-based Machine Translation, one of machine translations, which uses bilingual corpora as learning data. When trying to obtain enough quality of the translation by the Example-based Machine Translation just as it is, a large size of bilingual corpus of the community is needed. But when considering the cost, it is difficult to prepare a lot of parallel texts in all communities. So a method of using a small size of bilingual corpus to improve the quality of the Example-based Machine Translation by applying the transfer learning to this problem is supposed. And supposed method improves about 3 times of the translation quality by the score of automated evaluation. However when using only this method, the quality of the translation is decided only by bilingual corpus of the community prepared at the beginning, so the problem will still be left that the translation cannot keep pace with the changes of terms or phrased used in the community. To solve this problem, the system that parallel texts peculiar to a certain community are collected automatically while using customizing Example-based Machine Translation actually is constructed. This system improves the quality of the translation by automated enrichment of bilingual corpus of the community.

At last, an actual case of applying this system to a multilingual community is introduced, and the conclusion and the possibility of a support to a multilingual community by using the Example-based Machine Translation is described.

多言語コミュニティへの用例ベース翻訳の適用

森本 智史

内容梗概

本論文は、近年様々な現場で問題となってきた多言語コミュニケーションの支援について述べる。ここで言う多言語コミュニケーションとは、例えば医療現場における医師と外国人患者との間のコミュニケーションや、教育現場における教師と外国人の生徒やその親とのコミュニケーションの事である。それら多言語コミュニティでは、言語の壁の問題により母国語が違う人同士円滑なコミュニケーションが取れないなど、何かしらの支援が必要となっている。

言語の壁の問題に関しては、機械翻訳の技術が発達してきており、それらの技術を使う事により解消されつつある。しかし、実際に行われた機械翻訳のみを使った多言語コミュニケーションの実験により、多言語コミュニティ内で機械翻訳を使った際の特徴や問題点が明かになった。問題点の1つとして、多くの機械翻訳では、地域や団体といった特定のコミュニティや、医療現場や教育現場といった特定のドメインに依存するような場面で使おうとした場合、それらコミュニティやドメインでしか使用されない専門用語や言い回しなどが上手く伝えられないという事がある。すなわち、機械翻訳のみを使っただけでは円滑なコミュニケーションを支援しているとは言えない事になる。

上記の問題を解決するため、言語グリッドという言語インフラを構築する事を目標としたプロジェクトでは、これまで特定のコミュニティやドメインにおける多言語コミュニティ支援として、様々な取組みをしてきた。それは例えば、既存の機械翻訳に専門用語辞書や形態素解析などの言語処理を組み合わせる事で可能となった辞書組み込み翻訳であり、これにより、ドメインに依存するような専門用語の含まれる文章の翻訳の質を改善してきた。このように、言語グリッドでは、多言語コミュニティ支援を必要とするユーザが機械翻訳を上手く使えるような支援をしてきており、それらが実際様々な現場で利用されてきている。また、言語グリッドを用いた多言語コミュニティ支援のアプリケーションも多く開発されてきている。

以上のように、今まで言語グリッドを用いた多言語コミュニティ支援が多くなされてきており、そのいずれもある程度の実績を上げている。しかし、今後

更なる多言語コミュニティ支援を考えた時、支援を必要とする多言語コミュニティにカスタマイズされた機械翻訳が必要となってくると考えられる。これまで言語グリッドで多く使われてきた辞書組み込み翻訳などでは専門用語を含む文章の翻訳の質が改善されるが、辞書組み込み翻訳だけではコミュニティ独特の言い回しに対応できないなど、多言語コミュニティ支援にはまだいくつかの問題が残っている。本論文では、その問題の1つ、すなわち上で述べたコミュニティ独特の言い回しに対応するために、機械翻訳を多言語コミュニティ支援のためにどうカスタマイズして適応していくのかを議論する。

機械翻訳をコミュニティにカスタマイズする際には、機械翻訳の種類により手法が異なってくるが、一般的にそのコミュニティで使われる文章についての翻訳ルールや多言語の対訳コーパスが必要となってくる。しかし、コストの問題を考慮すると対訳コーパスを学習データとして用いた用例ベース翻訳という種類の機械翻訳を利用するのが適していると考えられるので、本論文では用例ベース翻訳に焦点を当てて議論する。ただ、用例ベース翻訳をそのまま用いて十分な翻訳の質を得ようとした場合、予めそのコミュニティの対訳コーパスを大規模に用意する必要があるが、コストの問題を考えると全てのコミュニティで十分な数の対訳を用意するのは困難であるという問題がある。そこで、機械学習の分野で研究されている転移学習の考え方を対訳コーパスに適応して、小規模な対訳コーパスを用意するだけでも翻訳の質を確保する手法を提案する。提案手法を用いる事により、翻訳機の自動評価の結果、自動評価に用いたスコアで約3倍の翻訳の改善が見られた。またこれだけだと、初期に用意するコミュニティの対訳コーパスだけで翻訳の質が決定してしまい、コミュニティで使う用語などが変化した場合に対応できないという問題が残ってしまう。この問題に対しては、カスタマイズした用例ベース翻訳をコミュニティで実際に使用していく中で、自然とコミュニティで使われる文章の対訳を収集できるようなシステムを考案し、構築した。構築したシステムでは、ユーザが意図する事なく対訳を自動的に収集する事により、コミュニティで使われる対訳コーパスが自動的に充実していき、利用している用例ベース翻訳の質が向上していく事となる。

最後に、上で構築したシステムを実際の多言語コミュニティの現場に適応した事例を紹介し、用例ベース翻訳を用いた多言語コミュニティ支援の成果や可能性について述べる。

Application of Example-Based Machine Translation to Multilingual Community

Contents

Chapter 1	Introduction	1
Chapter 2	Background	4
2.1	Intercultural Collaboration Experiment	5
2.2	Language Grid	7
Chapter 3	Multilingual Community Support	
	System using Language Grid	11
3.1	Language Grid Playground	11
3.2	Multilingual NOTA	14
3.3	Language Grid Toolbox	17
Chapter 4	Support to Multilingual Community by Example-Based Machine Translation	19
4.1	Outline of Example-Based Machine Translation	22
4.2	Improvement of Translation Quality by Example-Based Ma- chine Translation	23
4.3	Usage of Example-Based Machine Translation	24
4.4	Initial Evaluation of Example-Based Machine Translation	26
4.4.1	Evaluation Method	26
4.4.2	Evaluation	27
4.4.3	Conclusion of Initial Evaluation	28
Chapter 5	Multilingual Community Support System	33
5.1	Outline of System	33
5.2	System Architecture	34
5.3	Usage of Proposed System in Multilingual Community	39
5.4	Application of Proposed System to Multilingual Community	40
5.4.1	Purpose of Application	40
5.4.2	Application Method	41

5.4.3	Devisal on Application	43
5.4.4	Result of Application	46
Chapter 6	Conclusion	47
	Acknowledgments	49
	References	50

Chapter 1 Introduction

The opportunities for international exchange and the number of multicultural communities have increased in recent years as a result of globalization and the growth of the Internet. It is now essential to provide multilingual communication support in such places as hospitals and schools where foreign patients and students are full members of multilingual society. But in such places where multilingual communications are needed, the communication is often still difficult because of a problem of language barriers. And in such places, the major load in communications concentrates on multilinguals who are some members in the place because only they can interpret and translate. In the above example of hospitals countermeasures are taken that when foreign patient comes to have a medical examination the patient and a doctor have a communication by pointing parallel texts in a multilingual corpus prepared in advance. But when it is not enough a multilingual helps their communication by interpreting what they say. In schools, similar countermeasures such as pointing parallel texts in a multilingual corpus and interpreting by some teachers who can understand several languages are taken, too. That is to say a communication by pointing parallel texts costs a little to prepare but useful if enough, but if not enough the load of interpretations concentrates on multilinguals in each place. So it becomes essential that the environment where such a load does not concentrate on multilinguals and monolinguals can communicate in multilingual community without problems is supplied.

A machine translation is one of research results to the problem of language barriers that we cannot communicate with each other well in different languages like the above. It is expected that a machine translation is ideally able to translate sentences in a language into the other language automatically. But now much of the machine translation developed by the research of the machine translation is targeted to improve the quality of the translation of general sentences and there are not many researches of a machine translation that target to improve the translation quality of sentences that are often used in a certain community or domain. In such sentences that are often used in a certain com-

munity or domain generally include technical terms and peculiar expressions that are used only in the community or domain. And it is expected that when a machine translation cannot understand the peculiar rules used in such communities or domains it is difficult to ensure enough quality of the translation. That is to say, it is not always true that if a machine translation does not consider the community or domain and it is used as it is improvement of it does not conduce to expected quality of the translation when it translates sentences peculiar to a certain community or domain. In fact these phenomena appear conspicuously in case of a translation of sentences which include peculiar terms and expressions to a community or domain. There are many cases that a machine translation which can put up high quality of the translation of general sentences cannot put up enough translation quality of sentences including peculiar terms and expressions to a certain community or domain. Thus when using existing machine translation just as it is there are a lot of cases that it cannot support sufficiently to a certain multilingual community.

In this paper I describe supports to a multilingual community by using a machine translation provided so far at Chapter 2. As I described at the above, when using a machine translation just as it is there is possibilities that it cannot sufficiently support a certain multilingual community. Here I introduce an actual experiment of multilingual communications only using a machine translation, first. And I explain about an approach to the support to make sure that users can use a machine translation well and the problem of peculiar terms and expressions to a community or domain on the bases of the knowledge obtained through the experiment. Or I describe the framework in which we can easily create services which are able to support multilingual communities.

At Chapter 3, I show some case examples that supports of multilingual communities are provided in the framework described at the previous chapter. Here I show some applications that support a multilingual community and what kind of support has been provided by each application.

At Chapter 4, I show the Example-Based Machine Translation[6, 11] which is one kind of the machine translation. The Example-Based Machine Translation is the machine translation that improves the quality of the translation by

learning bilingual corpora but does not extract the rules about the translation from these corpora such as Statistical Machine Translation. It has the feature that it translates sentences by using parts of each parallel text in those corpora just as those are. So I discuss applying the Example-Based Machine Translation to a multilingual community because of the possibility that this feature can help multilingual communications well. When using the Example-Based Machine Translation in a community it is necessary to prepare a bilingual corpus including parallel texts of sentences used in the community, but it costs a lot and is difficult to prepare such bilingual corpus on a large scale. So I verify whether multilingual community support can be achieved not by using the Example-Based Machine Translation just as it is but by devising to prepare bilingual corpora used for learning.

I describe the method of applying the Example-Based Machine Translation inspected in Chapter 4 to actual multilingual communities at Chapter 5. Here I show the system I constructed for applying the Example-Based Machine Translation to a multilingual community. And I verify what kind of effect I can get or what kind of problems is still left when using the Example-Based Machine Translation in a multilingual community through the constructed system.

At last, I conclude this paper at Chapter 6.

Chapter 2 Background

The opportunities for multilingual communications have increased in recent years as a result of globalization. There are many cases that the communication on those is difficult because people in those can understand only their native tongue and no common language. For example, in a place of schools foreign students who need the Japanese guidance in public elementary, junior and senior high school in recent years tends to increase and reached more than 25,000 people at July in 2009¹⁾. In such a multilingual community the communication by pointing parallel texts on a multilingual corpus prepared for each community in advance is needed. But when enough communication is not had between those who are in the community by doing it, multilinguals who are in the community or invited to the community should help the communication. Interpretation by multilinguals is one of the solution of multilingual communications but it costs a lot and difficult to prepare enough number of interpreters in all multilingual communities and in all multilingual communications.

As the solution to the case that enough interpreters cannot be prepared like the above, there are also a lot of cases to try to communicate in multiple languages by an automatic translation such as the machine translation researched in the field of the natural language processing. That is the way of communications by making sentences in their native language, translating them into the other language which is the native language of the other person and conveying the translated sentences to the person.

In this chapter, it is shown that what kind of support has been performed to multilingual communities by a machine translation as related studies. The followings are introduced in this chapter: the experiment of multilingual communication supports by only using a machine translator and the knowledge obtained through it, the method how to apply a machine translation to a multilingual community based on the knowledge, and the multilingual service infrastructure that provides the environment where the method is easily used.

¹⁾ Referring to White Paper on Youth 2009 in Japan, published by the Cabinet Office

2.1 Intercultural Collaboration Experiment

Intercultural Collaboration Experiment[7] was held between Japan, China, Korea and Malaysia in 2002, which revealed what kind of phenomenon is observed and what kind of problem happens when people communicate each other in multilingual by using a machine translation. In this Intercultural Collaboration Experiment, the following task is established to make the above questions clear:

1. The goal of this experiment is to develop open source software by multinational teams
2. At least one college student from each country, which is Japan, China, Korea and Malaysia, participates in each team
3. The communication in the team is performed on bulletin board system (BBS) which has the translation function developed for this experiment and is called TransBBS, and all members of this experiment do not use any other communication tools except the TransBBS

The feature of the TransBBS is that when a user submits a message the process of the TransBBS goes with the following flow:

1. When a user submits a message, the user inputs sentences to convey in his/her native language first
2. The input sentences are translated into the other languages except the language of the original sentences on the TransBBS before submitting actually
3. The user confirms the displayed results of the machine translation and can translate original sentences again after modifying them if those results show the user that original sentences should be modified
4. The user repeats modification of original sentences until the result of the machine translation is convincing, and when it is judged that there are no problems with the translation results the user submit the message
5. The TransBBS supports users to communicate with each other in multilingual by saving not only original sentences but also translated sentences in multilingual and showing all of them to users

When a user submits a message all results of the translation in multilingual are saved like the above. The other users who can only understand a different language from the language of the original message can understand the user's

thought by seeing the translated message. Thus the mechanism that messages can be exchanged in multiple languages is built on the TransBBS used in the experiment.

The machine translator used in this Intercultural Collaboration Experiment is offered commercially and is made no improvements on itself. That is to say the purpose of this experiment is to verify what kind of phenomenon is obtained when people communicate in multilingual only by using the researched technology of a machine translation just as it is. The researches about a machine translation are developing step by step but there are still problems such as miscommunications often occur when a machine translation is used because of nonexistence of machine translate which can translate original message into the other language perfectly including the nuances of it. In the above experiment, in order to prevent to mistake communications by using a machine translation the following refinements were observed:

Self-initiated repair

Before submitting a message, a message contributor repeats repairing the original message to refine the translated results

Other-initiated repair

In order to share common knowledge between a message contributor and receivers, they repair errors of the translation collaboratively

For examples of self-initiated repair, the repeating refinement was observed that participants modify their original message by themselves by seeing and judging a result of the translation into a language what they were able to understand, most of them saw messages in English, if necessary. That is because they were able to confirm all results of the translation before submitting the message as the function of the TransBBS. One of the reasons why they conducted self-initiated repair is that the TransBBS, which was only a communication tool in the experiment, was a nonsynchronous communication tool and they desired to convey their meanings to the other participants as accurately as possible. In other words, self-initiated repair is the multilingual version of arrangement of sentences which is made for conveying correct meanings to the other. And this refinement is performed because a machine translation technology is not perfect

yet.

Other-initiated repair is the collaborative refinement between people that redeems the gap between the meanings that the contributor wants to convey and the meanings that the receivers receive caused by mistakes of a machine translation, which was not discovered by the above self-initiated repair. This phenomenon appears to have roots in a mistake of self-initiated repair because the contributor does not understand the language well which was used as the judgment of the translation result or there is a gap between the translation result in the language used as the judgment and in the other language, or so on. And to begin with this, when the contributor does not understand at all the language which a message is translated into, he/she cannot judge whether the translation result is correct or not and conduct self-initiated repair. When self-initiated repair does not work well as the above, receivers who see the translated message show points what they cannot understand or ask questions to the contributor in order to read the meanings. Such communication is also observed in monolingual communication but in multilingual communication to which the machine translation is introduced it is more difficult to communicate than in monolingual communication because of the problem of accuracy of a machine translation.

As the above, through the Intercultural Collaboration Experiment adopting of humans to machines was observed, such as understanding feature of the machine translation by repeating self-initiated repair and writing sentences for the machine translator to be able to translate correctly. That is to say, one result of this experiment was discovery of that it is necessary to help humans to write sentences which are able to be translated well by a machine translation for the support to a multilingual community. Additionally the problem was observed that technical terms and expressions are difficult to translate even if self-initiated repair is conducted.

2.2 Language Grid

A multilingual service infrastructure called the “Language Grid”[1], on which multilingual communication support is easily realized, has been developed based

on the Intercultural Collaboration Experiment in 2.1. In the framework of the Language Grid we can construct new services called language services for a multilingual community which requires some multilingual supports. Language services are realized by combining various language resources researched, developed and collected around the world. When building language services the Language Grid does not modify any language resources themselves but makes it easy to construct new language services by combining them appropriately. The Language Grid aims the realization of easy support to multilingual communication by constructing a language infrastructure on which such things is possible. That is to say, creating new language services necessary to multilingual communication on the Language Grid realizes easy multilingual communication support.

The Language Grid offers language resources as web services[8], and it is a realization method of the language infrastructure. Because language resources are able to be classified to some types such as the translation or the bilingual dictionary, it is designed so that the same type of language resource has the common interface for each type of the resource called standard interface. Therefore it is possible to combine those resources easily and it promotes to create new language services. That is, developers of language services need not to consider what language resources are actually available but can create new language services by assuming there are some kinds of abstract language resource such as the “Machine Translation” or the “Bilingual Dictionary”. And the Language Grid also constructs the framework in which a new language resource such as a bilingual dictionary or a bilingual corpus used in a certain community can be provided as a web service easily, so it is easy to share such language resources as used in a community and to combine those resources with the other language resources on the Language Grid.

Until now various language services have been created as a support to multilingual communities in the framework of the Language Grid. Examples of language services constructed are the Back Translation and the Translation with Dictionary.

The Back Translation is to translate sentences which are translated into a

certain language into the original language again. This was contrived to promote self-initiated repair observed in the Intercultural Collaboration Experiment at 2.1. In the Intercultural Collaboration Experiment participants judged whether the translation was correct or not by confirming the result of the translation in their non-native language. But it is impossible for persons to conduct self-initiated repair if they cannot understand the language into which the original message is translated and cannot judge the correctness of the translation. The Back translation enables us to judge in our native language whether we can convey our meanings in a message to others speaking different languages from ours even if we cannot understand their languages.

The Translation with Dictionary is the translation that enables to translate sentences which include peculiar terms to a certain community or domain correctly by using a bilingual dictionary of the community or domain while translating. The result of observation of the Intercultural Collaboration Experiment or the other cases shows that general machine translation system often delayed the communication because of its mistranslation of technical terms specialized in a community or domain. The Translation with Dictionary enables to convey messages correctly by what a bilingual dictionary including peculiar terms to a certain community or domain is prepared in advance and what those bilingual terms are used without passing through the machine translation system when those message are translated. For example, a general machine translation system translates a Japanese sentence “商と余りを計算しなさい。”, which meaning is “Calculate quotient and remainder.”, into an English sentence “Calculate trade and the rest.”. This translation result is incorrect and can not convey the original meanings. And so I prepare a bilingual dictionary which includes the following parallel texts in Japanese and English: “商” ↔ “quotient” and “余り” ↔ “remainder”. The Translation with Dictionary combine this dictionary translates the Japanese sentence into an English sentence “Calculate quotient and remainder.”.

The Language Grid has constructed the mechanism in which the support to multilingual communication can be achieved easily like the above. The point what we should pay attention is that the Language Grid makes no improvement

on language resources themselves. That is to say, the Language Grid enables to create the Back Translation process by connection two machine translation systems, which can help to learn the way to write sentences which is easy to translate by the machine translation system or to construct the Translation with Dictionary process by combining a machine translation with some dictionaries, which can help to convey technical terms in a community or domain correctly.

Usefulness of the Language Grid is cleared by what the number of joining to the Language Grid comes up to 119 as of January 2010. In the next chapter, I introduce several cases what kind of multilingual community support has been formed by using the Language Grid.

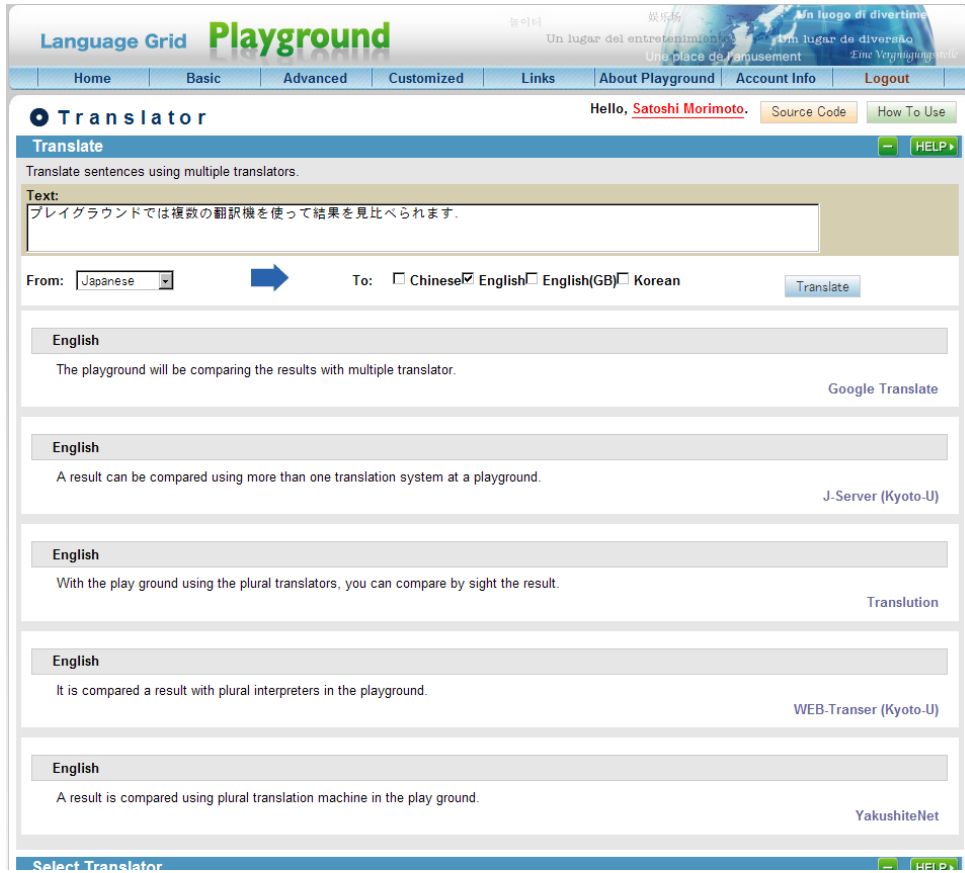


Figure 1: Translation Page in Language Grid Playground

Chapter 3 Multilingual Community Support System using Language Grid

Various multilingual community support systems have been constructed by using the Language Grid at 2.2. In this chapter I introduce some of those multilingual community systems.

3.1 Language Grid Playground

The Language Grid Playground¹⁾ (in the following I call this Playground) is the web application which can be used on the web browser (referring to Figure 1). The Playground offers several services that invoke available services in crossing way in view of each basic standard interface of the Language Grid and

¹⁾ <http://langrid.org/playground/>

services that are applicative services actually using language services created by combining several language resources. By the former, for example, it is possible to try to use language resources on the Language Grid by comparing results of the translation by some machine translation systems. And the main purpose of constructed the latter is to show samples such as what kind of services can be created by using the Language Grid actually.

One of the features of the Playground is that it has been built considering the support to developers who want to develop multilingual community support systems. On the Playground it is possible to try to use language services on the Language Grid actually as noted above. And the Playground also opens source codes composing the Playground at the same time. Those source codes help developers a lot who don't have much knowledge and techniques about web services.

There are several devices while constructing the Playground. One of those is to construct and accumulate well-used components called building blocks[10, 4]. These building blocks enable to create new language services easily on the Playground. Source codes of those building blocks are also open and this fact enables to provide the environment in which other developers of multilingual community support systems can construct new systems easily by using those building blocks.

Figure 2 illustrates the system's architecture of the Playground. Client side (web browser) accesses to the system of the Playground composed by the building blocks through the Ajax part. Those building blocks use some language resources and language services provided by the Language Grid. And the whole of the Playground system provides various language services.

The other feature of the Playground is to provide some multilingual community support services for a particular organization actually. For example, the Playground provides support services for Fujimi Junior High School and Kawasaki City Comprehensive Education Center in Japan. On those services functions required by such an organization are developed by using the Back Translation or the Translation with Dictionary and available.

The support system for Fujimi Junior High School supports communication

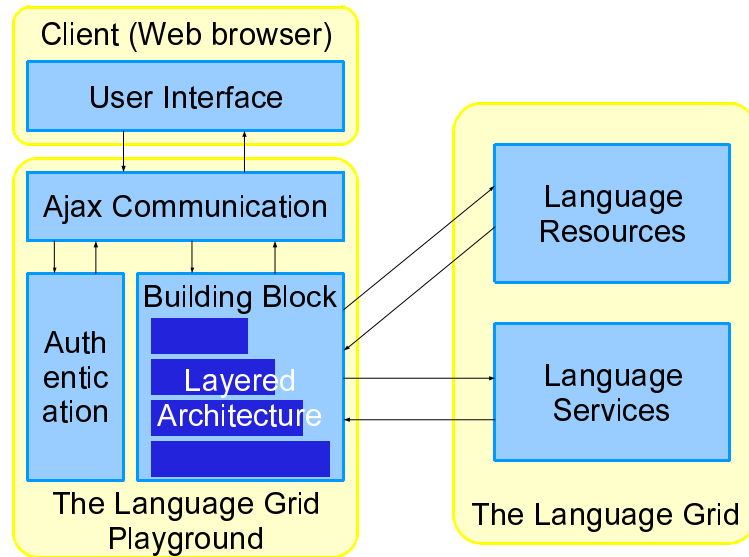


Figure 2: System Architecture of Language Grid Playground (drawn from [4])

between a teacher and a foreign student in the school. This support system supports face-to-face communication between them. Each of the teacher and the student inputs sentences they want to convey to the support system in their native language on one computer alternately. When translating those messages terms often used in the school are correctly conveyed by using the Translation with Dictionary of the school. And at the same time the result of the Back Translation is displayed on the system and enables for them to judge whether inputted messages are translated correctly in their native language. Moreover it is possible to refer to a multilingual corpus used in Fujimi Junior High School actually when they input messages. If they use some texts from the corpus parallel texts of those are used as the result of the translation, so they can convey meanings to the other without fault by using those templates. As the above, in this school more flexible multilingual communication is realized by using this support system otherwise pointing parallel texts.

Further, such support system for a specific organization has been developed by using the building blocks described above. We can realize to develop the support system at low cost by using the building blocks because of the reuse of common language processing such as the Back Translation and the Translation with Dictionary.



Figure 3: Multilingual NOTA displayed in Japanese and Korean

The Playground realizes the support to not only multilingual community but also developers of multilingual community support systems as the above. And in the support to multilingual community the Playground can provide appropriate functions to each community by combining several functions needed in the community.

3.2 Multilingual NOTA

The Multilingual NOTA (referring to Figure 3) is a collaboration tool extended a collaboration tool called NOTA¹⁾ to be able to use language services on the Language Grid. This Multilingual NOTA was constructed based on the NOTA which was used in several fields as a collaboration tool because I wanted to offer a place on which the Language Grid was used in an actual community. And the

¹⁾ <http://nota.jp/>

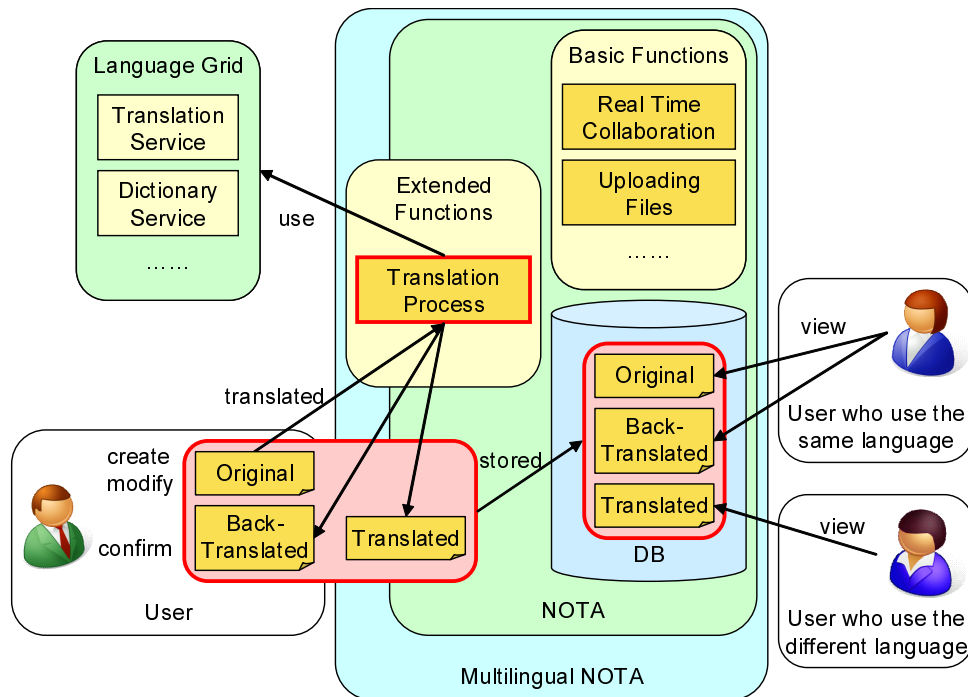


Figure 4: System Architecture of Multilingual NOTA

Multilingual NOTA has been used in the field of international exchange study contrived by Kubota and Kurokami laboratory, Kansai University, Japan, and greatly contributed to multilingual community in the study.

One of the features of the Multilingual NOTA is that collaboration nearly in real time, which is one of features of the NOTA, can be performed in multilingual. When users input messages on the Multilingual NOTA those messages are translated into the other languages decided in advance automatically by using translation services on the Language Grid. Each user can configure the language in which message are displayed, so on the Multilingual NOTA uses can collaborate in their native language because messages are displayed in the configured language. And also, the Multilingual NOTA shows the result of the Back Translation to not only the user who wrote the original message but also all other users by saving results of the Back Translation. This can help all users to learn the rules in order to write a message which has a tendency to be translated into other languages correctly from not only the self-initiated repair by the Back Translation but also those results.

Figure 4 shows system architecture of the Multilingual NOTA. The translation function of the Multilingual NOTA is achieved as one of extensions of the NOTA. And the translation function realizes the translation processing by using various languages services provided by the Language Grid. When a user writes a message he/she can check the adequacy of the translation by confirming the result of the Back Translation on the Multilingual NOTA. The other user who reads in a different language sees the translated message by the translation function. And a use who reads in the same language in which the original message was wrote can see not only the original message but also the result of the Back Translation and learn the way to write sentences which have tendency to be translated correctly from the others' way to write message.

This Multilingual NOTA has contributed to various fields of multilingual community such as international exchange study between Japan and Korea, and between Japan and Syria up to now [2]. The feature by the NOTA that image files can be easily displayed at free location of a screen has been used in multilingual a lot in the actual usage of the Multilingual NOTA.

For example, the exchange study between Japan and Korea shown in Figure 3 were held for the purpose that participants explain culture of their country and learn different culture each other by using pictures. The translation function of Multilingual NOTA was used in order to explain those pictures in multilingual. When a user wants to convey a message in multilingual on the Multilingual NOTA the user only inputs a message in his/her native language and so the message is translated into the other languages by the Translation with Dictionary service provided by the Language Grid automatically. It is necessary to register some terms needed in the exchange study in advance for bringing out the capability of the Translation with Dictionary, but if it is done once users of the Multilingual NOTA can communicate in multilingual with each other well. The Multilingual NOTA shows the result of the Back Translation as soon as a message is translated into the other languages as I described above. Thus, for example, a Japanese user who cannot read Korean messages can confirm whether the message translated into Korean has correct meanings what he/she wants to convey to a Korean user from the result of the Back Translation in



Figure 5: BBS on Language Grid Toolbox

Japanese.

The Multilingual NOTA, which has the features that the interface of it is intuitive and easy to understand, and it supports multilingual community by using language services on the Language Grid, has also contributed greatly to some multilingual communities such as international exchange study between Japan and China, and between Japan and Syria. I feel certain that this Multilingual NOTA shows that support by using the Language Grid can actually contribute to multilingual community.

3.3 Language Grid Toolbox

Figure 5 is the screenshot of one service on the Language Grid Toolbox¹⁾ (in the following I call this Toolbox). The Toolbox provides a series of tools for

¹⁾ <http://langrid-tool.nict.go.jp/toolbox/>

supporting communication in multilingual communities by using the Language Grid, which is developed by the National Institute of Information and Communications Technology (NICT), Japan (cited from the web page). The Toolbox is based on the Playground at 3.1 and has been developed by focusing on functions required in actual multilingual communities. The Toolbox is desired to be actually used in a lot of multilingual communities.

The Toolbox achieves multilingual community support by taking functions in it from the Playground and making new functions, which are necessary to multilingual communities. The Playground has contributed greatly in terms of supporting developers of multilingual community support systems and letting users to understand deeply about the Language Grid. But it is too complicated to use in a multilingual community actually because of many extra functions. So the Toolbox achieves to be used actually in multilingual communities by being pruned those extra functions not needed in multilingual Communications and being attached necessary functions.

Multilingual BBS is one of functions for supporting multilingual communities on the Toolbox. It has several same functions as the TransBBS used in the Intercultural Collaboration Experiment at 2.1, which provides the automatic translation and chance to confirm the translation result before posting messages. And as improved point, the following functions are attached to the Toolbox:

1. The Translation with Dictionary is used as the translation
2. The Back Translation is also available while checking the result of the translation

In addition to the BBS picked up as the example, functions considering accumulated knowledge obtained from the Intercultural Collaboration Experiment and actual usage of the Language Grid are attached to the Toolbox and so make it to a tool for supporting more smooth multilingual Communications.

Original Sentence

それと久しぶりに、初めてビスケットをする子に教える機会もありました。(This means “And after a long interval there was a chance to teach a child who plays with the Viscuit for the first time (how to play with the Viscuit).”)

The Result of Machine Translation

And there was also a chance to tell the child who makes the biscuit after a long time for the first time.

The Result of Translation with Dictionary

And there was also a chance to tell the child who makes Viscuit after a long time for the first time.

What is expected

A part of the original sentence wants to be translated into not “makes Viscuit” but “plays with the Viscuit”.

Figure 6: Limitation of Translation with Dictionary

Chapter 4 Support to Multilingual Community by Example-Based Machine Translation

Multilingual community support by using a machine translation has been improved gradually by being innovated new functions really needed as I introduced at Chapter 2 and Chapter 3. And until now the support to multilingual communities is achieved by combining several language resources as they are. For example, those are the Back Translation, which helps to conduct self-initiated repair, and the Translation with Dictionary, which is constructed by combining a machine translation with some bilingual dictionaries.

But it was found out that the Translation with Dictionary cannot solve all the problem of the mistranslation in a multilingual community from the observation of activities performed in several multilingual communities. For example, Figure 6 shows that the Translation with Dictionary cannot convey

a nuance included in the original message in Japanese after translating it into English. Here the original sentence in this figure is extracted from a community site managed by NPO Pangaea¹⁾, which is one of organizations and uses the Language Grid in actual activities.

In this example the word “ビスケット” is translated not into “biscuit” but into “Viscuit”²⁾ correctly by the Translation with Dictionary³⁾. But this is the limitation of the Translation with Dictionary; the Translation with Dictionary can only replace specific terms. So it has been clear from experiences of supporting various multilingual communities that it is difficult for only the Translation with Dictionary to translate a part of the original message not into the result “makes Viscuit” but into “plays with the Viscuit” correctly.

In order to solve the problem such as above the following two ways to support multilingual community by using a machine translation are possible:

1. To support interaction between humans or between humans and machines
2. To customize machine translation for a multilingual community

1 is the idea that to help self-initiated repair such as the Back Translation causes for humans to adapt to machines and for a machine translation to translate messages more accurately. Or to repeat other-initiated repair only monolinguals can collaboratively translate a document into the other language by using a machine translation or the Back Translation such as the Collaborative Translation[5]. In the idea of 1, a machine translation is consistently a tool to communicate and humans can communicate smoothly in multilingual by using it with some devices.

On the other hand 2 is to apply machine translations researched widely in the field of the Natural Language Processing to multilingual communities by customizing them to each community. In recent years the technology of a machine translation is improving as the result of researches but there are a few movements to optimize machine translation technologies to a multilingual

¹⁾ <http://www.pangaeaan.org>

²⁾ <http://www.viscuit.com/>

³⁾ “ビスケット” normally means “biscuit”, but almost means “Viscuit” in the case of NPO Pangaea

community which really needs them in order to translate sentences including peculiar expressions to the community accurately. So in fact the phenomenon has occurred that the quality of the translation of peculiar sentences to a community by a machine translation system is low and it is judged not to be working even if it can translate general sentences well. One of solutions to these problems is 2, and the purpose is to customize and optimize a machine translation to each multilingual community to translate peculiar expressions used in the community correctly. It is expected that the example problem in Figure 6, which is not solved by the Translation with Dictionary, is solved as the result of those customizations and optimizations.

This paper targets to introduce the idea of 1 to the idea of 2. As far as the idea of 2, however, I have no idea to improve the algorithm of machine translate itself. But I leverage the machine translate that improve the translation quality by learning bilingual corpora and want to achieve to customize it to a certain community or domain by preparing suitable corpora for the community or domain and to realize the multilingual community support. I also target to build the framework in which a bilingual corpus is enriched by the interaction between humans or between humans and machines such as 1.

As the underlying machine translation customized to a multilingual community I decided to use the KyotoEBMT developed in Kurohashi laboratory, Kyoto University, Japan[3], which is one of the kinds of the Example-Based Machine Translation. The detailed reason why I decided to use the KyotoEBMT is described below, but the Example-Based Machine Translation can ensure high quality of the translation of specific sentences to a certain domain when it learned a bilingual corpus of the domain even if it is small sized. The bilingual corpus including parallel texts often used in a community can be prepared only in the community so it costs a lot to prepare in a large scale. Therefore the above feature what the Example-Based Machine Translation has is very useful to be applied to various multilingual communities.

In this chapter I suggest how to apply the Example-Based Machine Translation to a multilingual community after describing the outline of it. And I examine whether the usage of the Example-Based Machine Translation in the

suggested method really helps multilingual communities by some primary evaluations.

4.1 Outline of Example-Based Machine Translation

There are some kinds of the machine translation researched in the field of the Natural Language Processing. These are the Rule-based Machine Translation, the Statistical Machine Translation and the Example-Based Machine Translation, and each feature is as follows:

Rule-based Machine Translation

Rules about the translation are described by humans and used while translating. In order to improve the quality of the translation those rules should be enriched in a large scale.

Statistical Machine Translation

A bilingual corpus is prepared and rules about the translation are extracted from it by machine learning. A large scale of the bilingual corpus is needed for extracting those rules.

Example-Based Machine Translation

A bilingual corpus is prepared and parallel texts in it are kept up as learning data. When translating, the Example-Based Machine Translation system search examples which are similar to the original sentence from those parallel texts in the corpus. Repeating this action and combining those texts coming from the actions make a sentence as the translation result.

I think the Example-Based Machine Translation is the best when I select the most suitable machine translation to this research from above kinds of the machine translation. Because the two kinds of the machine translation expect the Example-Based Machine Translation cost very much when the quality of the translation is kept high in a certain community. About the Rule-based Machine Translation rules of the translation which is about expressions peculiar to the community should be made. But only experts about translation rules can make those rules and it cost a lot for them to make those rules which covers all expressions used in the community. And the Statistical Machine Translation needs a large-scaled bilingual corpus including many expressions used in the

community in order to extract translation rules about sentences in the community. But there are no or few multilinguals in the multilingual community which needs some supports about language barriers, so it is difficult for them to create a bilingual corpus of the community in a large scale. In contrast the Example-Based Machine Translation, which I use in this research, has the feature that high quality of translating sentences which domain is similar to the learned bilingual corpus can be achieved even if the corpus is comparatively in a small scale. This feature is assumed to be useful to a certain community or domain in which much cost cannot be paid to create translation rules or a bilingual corpus of the community or domain, so I decided to use the Example-Based Machine Translation in this research.

The detailed behavior of the KyotoEBMT, the Example-Based Machine Translation used in this research, is the following. The KyotoEBMT makes tree structures of a sentence by applying as much available language resources such as the morphological analyzer and the dependency parser as possible to a bilingual corpus to be learned and aligns each subtree in two languages. When translating, the KyotoEBMT analyzes an input sentence to make a tree structure by those available languages resources. After this analysis the KyotoEBMT searches subtrees included in examples in the corpus which is similar to a subtree included in the tree structure of the input sentence, selects the subtree that is the largest size in them and get the subtree in the other language aligned to it. Repeating those actions makes a translation result by processing all subtrees in the tree structure of the input sentence. That is to say, the KyotoEBMT does not learn rules of the translation from a bilingual corpus but uses examples in the corpus as they are, so this fact makes the quality of the translation of sentences which domain is similar to the learned corpus high.

4.2 Improvement of Translation Quality by Example-Based Machine Translation

The support to a multilingual community by using a machine translation has been mainly to encourage self-initiated repair by the Back Translation and to translate specific terms well by the Translation with Dictionary as I described

at 2.2 and Chapter 3. Especially the latter is one of successful solutions to the problem that messages including peculiar terms are not translated well by a general machine translation and is almost used in the support to a multilingual community by using the Language Grid. But now more effective solution is needed to the problem that the Translation with Dictionary cannot translate sentences including peculiar expressions to a community well as I described at the beginning of this chapter.

So in this paper I suppose the support to a multilingual community by using the Example-Based Machine Translation. For example, I apply the Example-Based Machine Translation to the problem of Figure 6 at the beginning of this chapter. This is that I prepare a bilingual corpus of NPO Pangaea including a parallel text “ビスケットをする” \leftrightarrow “play with the Viscuit”. And this corpus will make the quality of the translation by the Example-Based Machine Translation high.

The Example-Based Machine Translation is used to tackle not only specific terms but also peculiar expressions to a community and to achieve more available support to a multilingual community. To translate sentences by the Example-Based Machine Translation needs a bilingual corpus including parallel texts often used in a community or domain but it costs a lot and is difficult to prepare such a bilingual corpus in a large scale. So I devised to ensure the quality of the translation by the method of the next section even if a bilingual corpus is prepared in a small scale.

4.3 Usage of Example-Based Machine Translation

The Example-Based Machine Translation has the feature that a bilingual corpus of a community realizes high quality of the translation of sentences used in the community as described above. But when it is used in the support to a multilingual community, a bilingual corpus of the community (in the following I call this community corpus) should be prepared and enriched in a large scale to cover whole sentences used in the community. It is, however, costs very much and impossible to create such a large-scaled community corpus in every multilingual community. Or there is also a problem that when only small-scaled

community corpus is prepared as a result of reducing the cost the quality of the translation is not ensured.

So I suppose that the idea of the Transfer Learning is applied to the problem of the community corpus and the quality of the translation is ensured at low cost.

The Transfer Learning is defined as “the problem of retaining and applying the knowledge learned in one or more tasks to efficiently develop an effective hypothesis for a new task”. That is the method used in the field of the Machine Learning that the knowledge learned from a domain is transferred to new knowledge of the other domain. In the field of the Transfer Learning the knowledge already learned is called a source domain and new knowledge is called a target domain.

In the proposal of this paper the target domain is a certain community corpus. That is new knowledge that is wanted to acquire is the knowledge about parallel texts used in a certain community or domain.

On the other hand the source domain is a general bilingual corpus (in the following I call this general corpus). There have been a large number of researches that use a large-scaled bilingual corpus and such bilingual corpus is generally domain-independent. That is to say learning such a large-scaled general corpus which is easy to acquire enables to get the knowledge about parallel texts which are domain-independent. And this knowledge is the knowledge of the source domain itself.

In order to apply the proposed method actually large-scaled general corpora and small-scaled community corpora are needed to be learned by the Example-Based Machine Translation and to be used when it translates sentences at the same time.

The reason why this can achieve the Transfer Learning is that the Example-Based Machine Translation has the feature that it translates a sentence by using parallel texts included in learned bilingual corpora which structures are similar to that of the input sentence. The grammar or the structure of specific sentences to a community is rarely out of those of sentences in general corpora completely. That is, the knowledge of a source domain from large-scaled general

corpora can be used well to the domain-independent part such as that. And the domain-specific part such as peculiar terms and expressions to a community is translated well by using the knowledge from the small-scaled community corpus if it has at least one parallel text including such terms or expressions. Because such terms and expressions are rarely included in a general corpus and when examples which have a similar structure to them are searched from learned corpora found examples are from the community corpora with high probability.

As described above to learn from a large-scaled general corpus as the knowledge of the source domain and a small-scaled community corpus as the knowledge of the target domain at the same enables to acquire the large-scaled community corpus.

In the next section as an initial evaluation I verify how the quality of the translation by using the method proposed in this section changes when the Example-Based Machine Translation is used to translate sentences of a certain community or domain actually.

4.4 Initial Evaluation of Example-Based Machine Translation

4.4.1 Evaluation Method

The Example-Based Machine Translation was evaluated about that it translates sentences from Japanese into English by using a large-scaled general corpus as a source domain and a small-scaled community corpus as a target domain. The KyotoEBMT of Kurohashi laboratory in Kyoto University, Japan was used as an Example-Based Machine Translation. And a bilingual corpus extracted from abstracts of papers (in the following I call this corpus papers abstract corpus) was used as a large-scaled general corpus. I used this papers abstract corpus because it contained the most number of parallel texts in available corpora and was often used in evaluating the KyotoEBMT. On the other hand a bilingual corpus about dialogues in a hospital reception (in the following I call this corpus hospital reception corpus) was prepared as a small-scaled community corpus. Parallel texts in this hospital reception corpus were extracted from parallel texts actually used in Communications between doctors or nurses and foreign

patients in some hospitals and bilingual documents such as bilingual booklets. The papers abstract corpus contains about one million pairs of parallel texts and the hospital reception corpus 2101 pairs of parallel texts.

As the test sentences for the evaluation of the Example-Based Machine Translation by the KyotoEBMT 211 pairs which are ten percent of the number of the hospital reception corpus were picked up. And as learning data several types were prepared, which were the data learning the papers abstract corpus and various numbers of pairs of parallel texts picked up from the hospital reception corpus. The number of pairs from the hospital corpus was changed every 210 from 210 to 1890, which was not used as test sentences. That is to say ten types of learning data, which is the papers abstract corpus added every 210 pairs from 0 pair to 1890 pairs from the hospital reception corpus, were prepared.

4.4.2 Evaluation

One of the automatic evaluation indicators of machine translation systems is the BLEU[9]. The BLEU evaluates a machine translation by comparing a sentence translated by it with an answer sentence and considering the number of overlaps of the sequence of words, which is called N-gram. It is often used as a de-facto standard evaluation indicator in the field of the research of a machine translation because it is well known that there is a strong correlation between the score of the BLEU and human evaluations about adequacy and fluency.

When the score of the BLEU was calculated, five times of trial experiments were conducted by randomizing the order of parallel texts in the hospital reception corpus. In each trial translations from Japanese into English by the KyotoEBMT were made by using learning data which were prepared by changing the additional number of parallel texts from the hospital reception corpus in the way as described above. Figure 7 is the result of those trials.

Figure 7 shows that more and more numbers of parallel texts added from the hospital reception corpus tend to make the quality of the translation higher and higher. Especially the change from 0 pairs to 210 pairs, that is the change of learning data from only using the papers abstract corpus to using the papers abstract corpus and a little from the hospital reception corpus, is seen to be

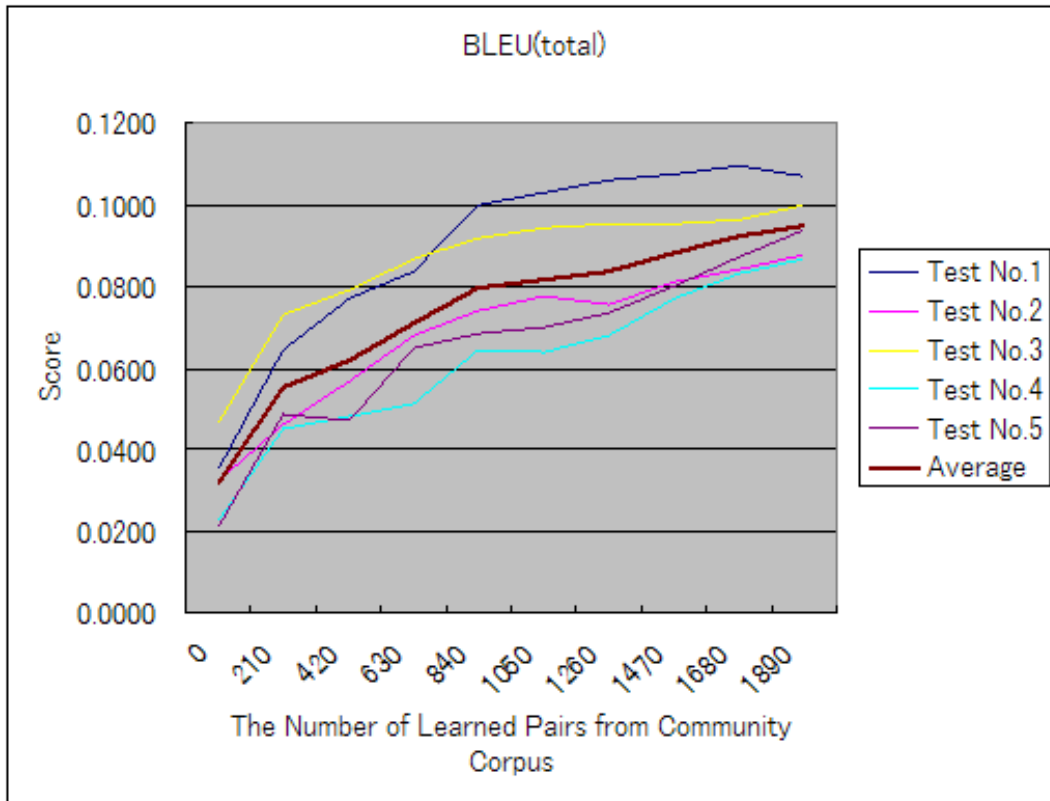


Figure 7: Automatic Evaluation by BLEU

the biggest in those changes so this figure says that a little scaled community corpus can improve the quality of the translation.

4.4.3 Conclusion of Initial Evaluation

In order to back up the evaluation of 4.4.2 human evaluation of the Example-Based Machine Translation from English to Japanese was also done by changing the learning number from the hospital reception corpus to 0 pair, 210 pairs, 1050 pairs and 1890 pairs. As the evaluation two indicators are used; adequacy, which shows how much of the meaning expressed in the gold-standard translation is also expressed in the target translation, and fluency, which shows how the fluency of the translation is judged. In addition the Google Translate¹⁾, which is one of the Statistical Machine Translation, was evaluated at the same time by way of comparison. Figure 8 and Figure 9 show results of this evaluation. The average μ and the standard deviation σ is as the followings: adequacy is

¹⁾ <http://translate.google.com/>

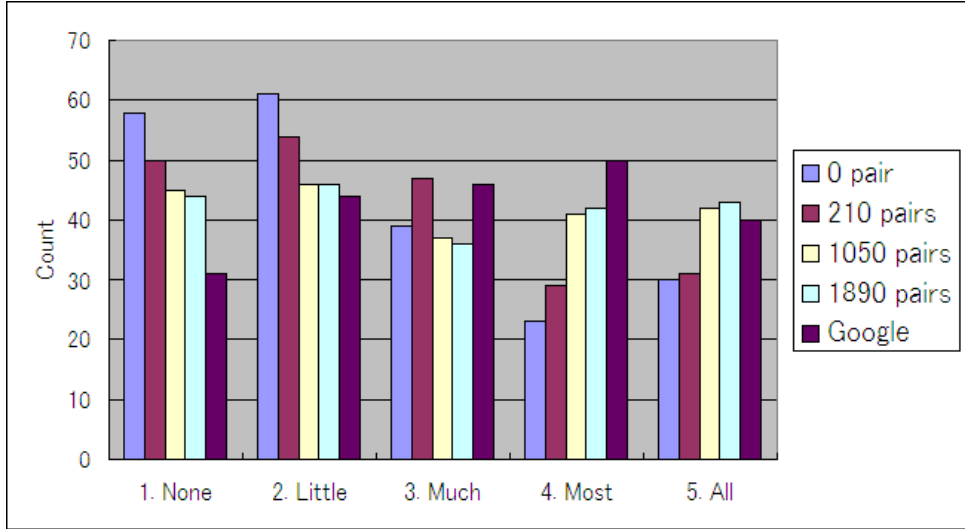


Figure 8: Distribution of Adequacy

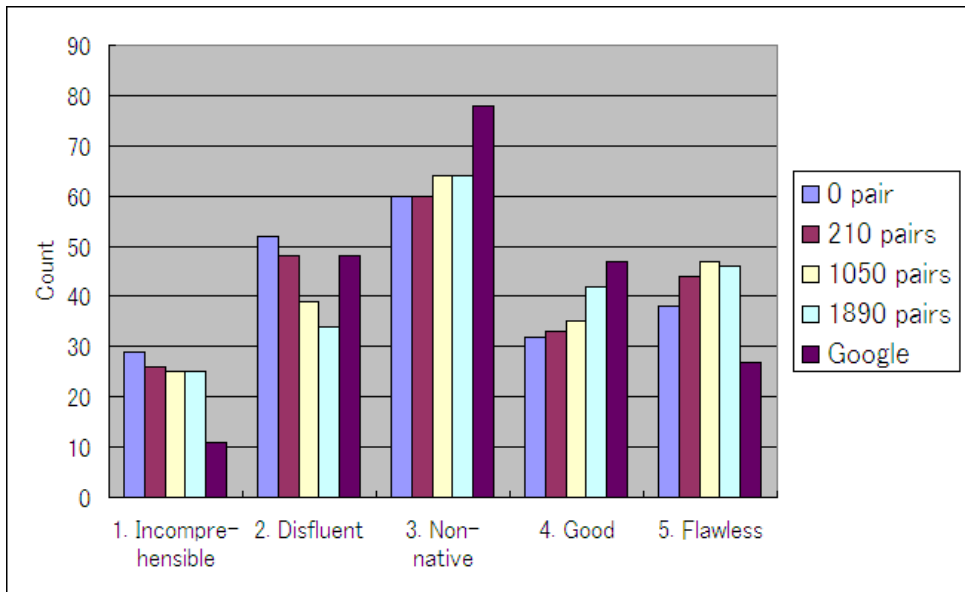


Figure 9: Distribution of Fluency

(0 pair) $\mu = 2.55, \sigma = 1.37$, (210 pairs) $\mu = 2.70, \sigma = 1.36$, (1050 pairs) $\mu = 2.95, \sigma = 1.44$, (1890 pairs) $\mu = 2.97, \sigma = 1.44$ and (Google) $\mu = 3.11, \sigma = 1.34$, and fluency is (0 pair) $\mu = 2.99, \sigma = 1.29$, (210 pairs) $\mu = 3.10, \sigma = 1.31$, (1050 pair) $\mu = 3.19, \sigma = 1.30$, (1890 pairs) $\mu = 3.24, \sigma = 1.29$, (Google) $\mu = 3.15, \sigma = 1.07$. These results show that increasing the number of parallel texts from the hospital reception corpus improves the quality of the translation.

And as the detailed validation of changing the score of the BLEU I examined which part of 211 sentences used in the test of 4.4.2 were improved or worsened by comparing the result of 0 pair, 1050 pairs and 1890 pairs. The blanket impressions of the result are the following:

1. The usage of examples at the level of words and phrased raises the score of the BLEU
2. More and more number of parallel texts from the hospital reception corpus are learned, the quality of the translation becomes higher and higher because of the increase of examples similar to some of sentences for the test
3. Extra parts which are near by specific terms or phrased are sometimes used in the translation and this fact vitiates the quality

1 is the case example that to add a community corpus to a general corpus improves the quality of the translation. That is, transferring the knowledge of a source domain, that is a general corpus, to the knowledge of a target domain, that is a community corpus, is succeeded by only a little scaled community corpus. An example about 1 is the following.

Sentences used in the test include a sentence of “過激な運動は避けて下さい。” in Japanese, which parallel translation is “Please avoid hard exercise.” in English. The result of the translation by the KyotoEBMT using the learning data of 0 pair is “Please stop and for a extreme exercise .”, and it is wrong. On the other hand the KyotoEBMT of 1890 pairs can translate it into “Please avoid a extreme exercise .”, and this tells that the quality of the translation is improved. The reason of the improvement is there is a parallel text in learned 1890 parallel texts from the hospital reception corpus; “眠くなりますので、車の運転や危険な作業は避けてください。” ↔ “Please avoid driving and dangerous work because this medicine will make you sleepy”. The adequate usage of the part of “...は避けてください” ↔ “Please avoid ...” in it changes the part of “Please stop and for”, which is not translated well by the KyotoEBMT of 0 pair, to be translated correctly.

2 shows that the enrichment of a community corpus as a target domain raises the quality of the transfer and the translation. Or the proposed method using the idea of the Transfer Learning effectively improves the quality of the

translation and more and more knowledge of the target domain are more and more effective. Therefore it is also needed to enrich a community corpus at low cost not only to apply the proposed method to the Example-Based Machine Translation when it is actually used in a community. The following is the example of this fact.

Sentences used in the test include a sentence of “胸が焼きこがすように痛い” in Japanese, which parallel translation is “I have a burning pain in my chest” in English. The result of the translation by the KyotoEBMT of 0 pair is “My chest hurts like baked singe .”. On the other hand the KyotoEBMT of 1050 pairs can translate it into “I have a burning pain in my chest .”, which quality is very high. The reason of this phenomenon is there is a very similar parallel text to the tested sentence in learned 1050 parallel texts from the hospital reception corpus; “胸が焼き焦がすように痛いです。”↔“I have a burning pain in my chest.”. When the KyotoEBMT searched examples similar to the tested sentence from the learned parallel texts is was judged to be most suitable to select. This is an extreme example but sentences often used in a certain community include peculiar terms and expressions to the community so the phenomenon such as this will sometimes occur. Therefore the enrichment of a community corpus enables to raise the quality of the translation.

3 is that the enrichment of a community corpus causes the reduction of the quality of the translation. The example of this phenomenon is the following.

Sentences used in the test include a sentence of “爪に異常があります。” in Japanese, which parallel translation is “My fingernails are abnormal.” in English. The result of the translation by the KyotoEBMT of 0 pair is “My nails i is unusual .” and the translation is not perfect because it includes the extra word “i” and it uses “is” in spite of the plural subject but it can be barely read the meaning. On the other hand the KyotoEBMT of 1890 pairs translates it into “I have trouble seeing my nails .”. That is the insertion of the extra word “seeing” changes the meaning. The reason of such insertion is the following: that is because there is a parallel text in learned parallel texts from the hospital reception corpus; “視覚に異常があります。”↔“I have trouble seeing.”. In this parallel text it is desired that the part of “...に異常がありま

す”↔“I have trouble . . .” is used when being translated but the part of “I have trouble seeing . . .” including “seeing” was actually used.

With regard to this problem it is possible to solve it by the devisal to enrich a community corpus like as 2. For example of the above problem, it has possibility to improve the translation that a parallel text including the alignment of “視覚”↔“seeing” is learned or parallel texts including the alignment of “. . . に異常があります”↔“I have trouble . . .” are learned more. But whether learned community corpus should be enriched naively or not is needed to be checked up. And it also remains to be solved that what kinds and how much of parallel texts should be prepared in order to improve the quality of the translation.

Chapter 5 Multilingual Community Support System

It is clear as desired at Chapter 4 that the KyotoEBMT learning a large-scaled general corpus and a small-scaled community corpus can support the multilingual community. A community corpus is prepared in a larger and larger scale; it can translate sentences in the community more and more correctly. But in fact it costs and is difficult to prepare a community corpus for a multilingual community which requires some supports in a large scale in advance.

So I constructed a system using the KyotoEBMT customized to a community that firstly the quality of the translation of it is not high but to use it more and more in the community can raise the quality of translation more and more and at some day in the future it becomes to be specialized to the community. On this system a community corpus to be prepared can be in a small scale at first but the usage of it in a community enables to enrich the community corpus automatically and as a result the translation of the customized KyotoEBMT is improved.

5.1 Outline of System

Figure 10 illustrates the feature of the constructed system. On this system translated sentences into a language by the KyotoEBMT can be modified by native or near native speakers. The envisaged case is that some sentences translated by the KyotoEBMT whose meanings are almost clear but need to be modified by a few users. Or regarding translated sentences whose meanings are not clear at all bilinguals or multilinguals can also modify them. The modified sentence is coupled the original sentence and the pair is added to a community corpus. After this when a sentence is translated by the KyotoEBMT the new community corpus is used and as a result the quality of the translation is expected to be improved. That is to say by using this system the KyotoEBMT is gradually customized to a community as a result of the enrichment of the community corpus. And the support to multilingual community by the Example-Based Machine Translated can be achieved by applying this system to an actual

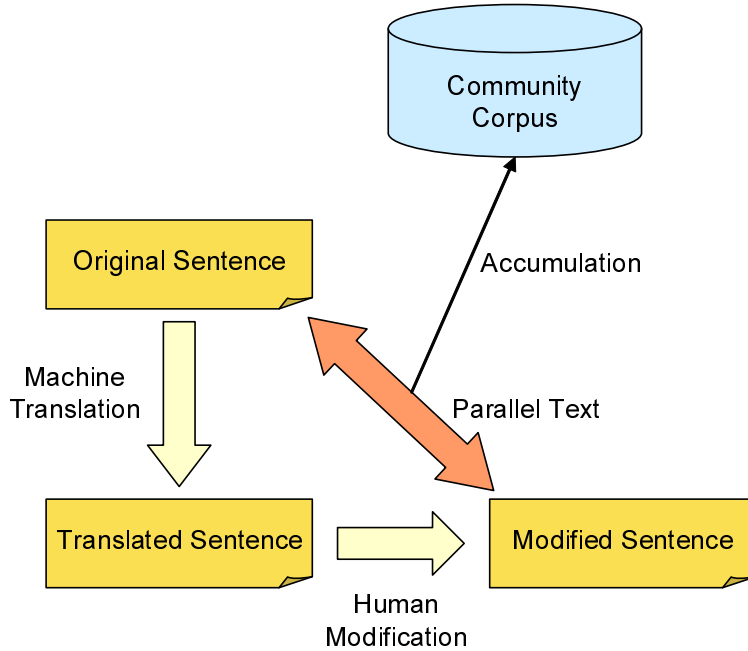


Figure 10: Automatic Enrichment by Modification of Translated Sentence

multilingual community.

5.2 System Architecture

This system is based on the Toolbox introduced in 3.3.

The Toolbox has various types of support systems for a multilingual community and one of them is the BBS. In the BBS users can read discussions in their native language as described in 3.3. Or users can use translations to other languages and confirm the results of the Back Translation when posting a message.

Modification of translated sentences is another function of the BBS on the Toolbox. Translated messages by the translation function sometimes cost a few to understand their meanings because they have almost clear meanings but are non-native. The function of modification of a translated sentence is that users whose native language is the same with the one of the translated sentence or those who can understand deeply about the language can modify the translated sentence. This function of modification of a translated sentence can help the other users to understand it and to promote smooth multilingual communica-

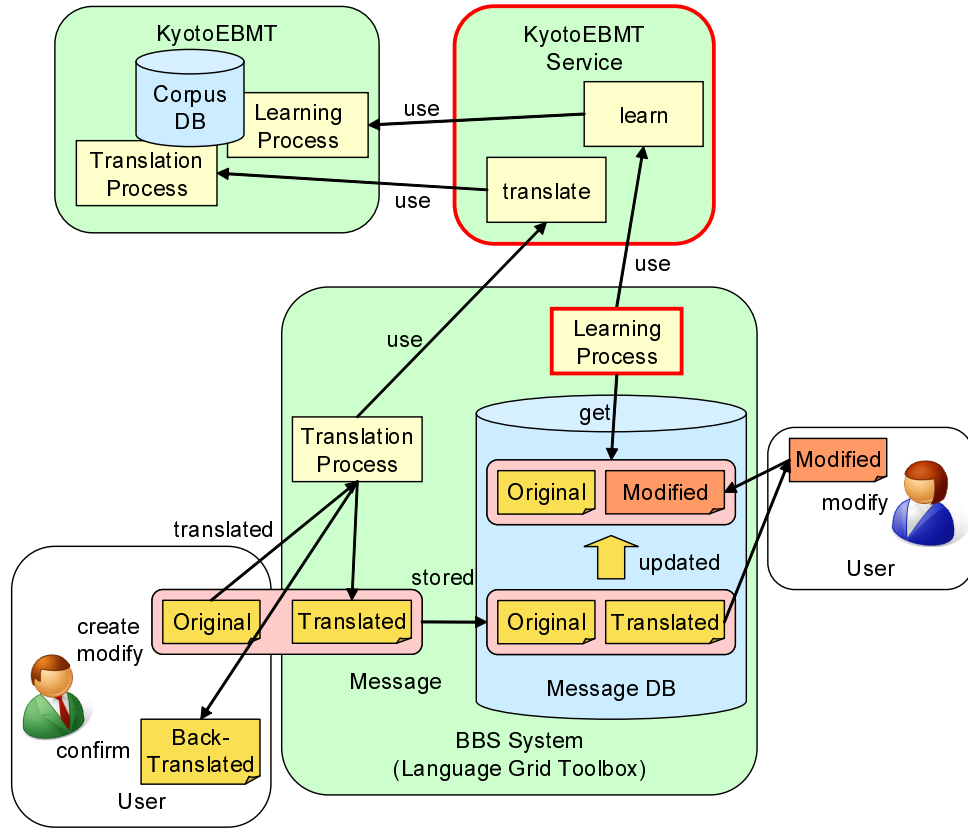


Figure 11: System Architecture

tions in the community. Or bilinguals or multilinguals in the community also help the other users by the function.

The modification of a translated sentence is actually made in the community site of NPO Pangaea described above. On the community site of NPO Pangaea the function of the modification of a sentence enables users to communicate with others in their native language and to share information easily. Because of such actual fact, the Toolbox also innovates this function into the BBS and realizes the multilingual community support.

The proposed system using the Toolbox is shown in Figure 11. And Figure 12 shows the class diagram of the main functions of the system. The KyotoEBMT-Service in those figures is a wrapper program created with the purpose of using the KyotoEBMT as a web service. I have to build the wrapper program in order to use it as a web service because it is just a program. This method of creating a wrapper program is often leveraged in the Language Grid. The

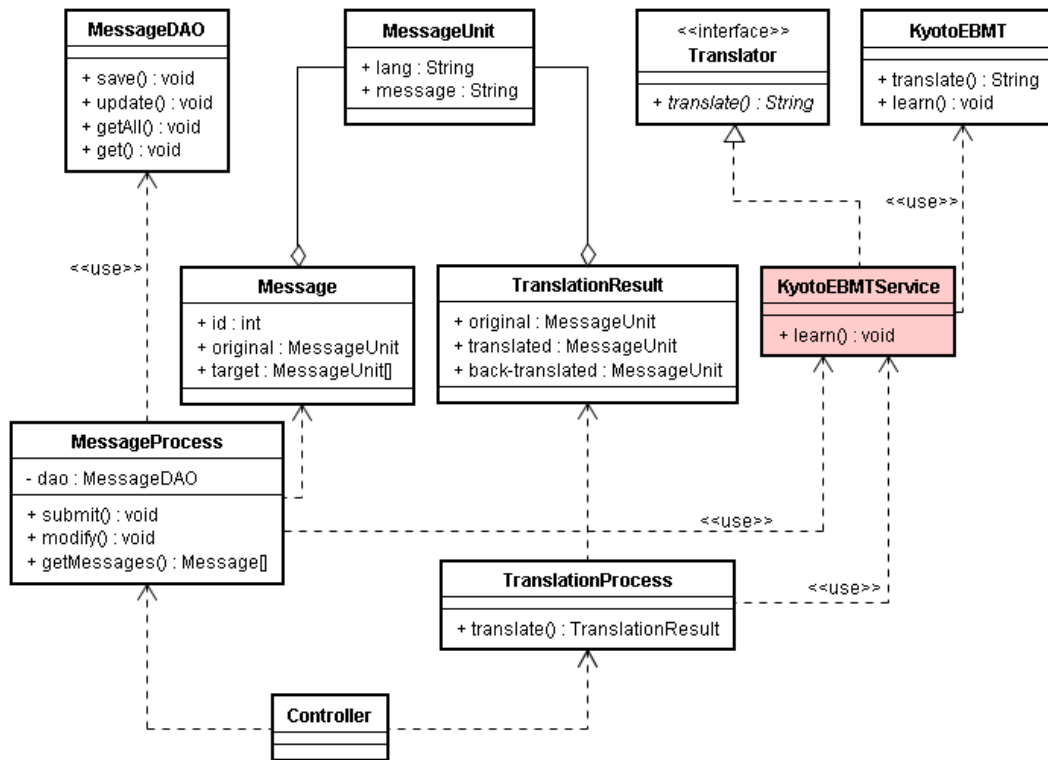


Figure 12: Class Diagram of System

KyotoEBMTService is also created with the interface of the Translator, which is one of the standard interfaces used in the Language Grid, and can be easily used from the Toolbox which uses translation services on the Language Grid.

When a user posts a message this system will save the message to its database as illustrated in Figure 13. The detailed flow of this action is listed as follows and a pair of the original message and the translated message will be stored:

1. A message wrote by a user is translated firstly before posted. The translation process uses the KyotoEBMT which is customized to a community which he/she belongs to and is expected to translate the sentence in high quality.
2. He/she repeats to modify the original message by confirming the result of the Back Translation. Such self-initiated repair can prevent a miscommunication between him/her and other users.
3. After his/her modifying he/she posts it.

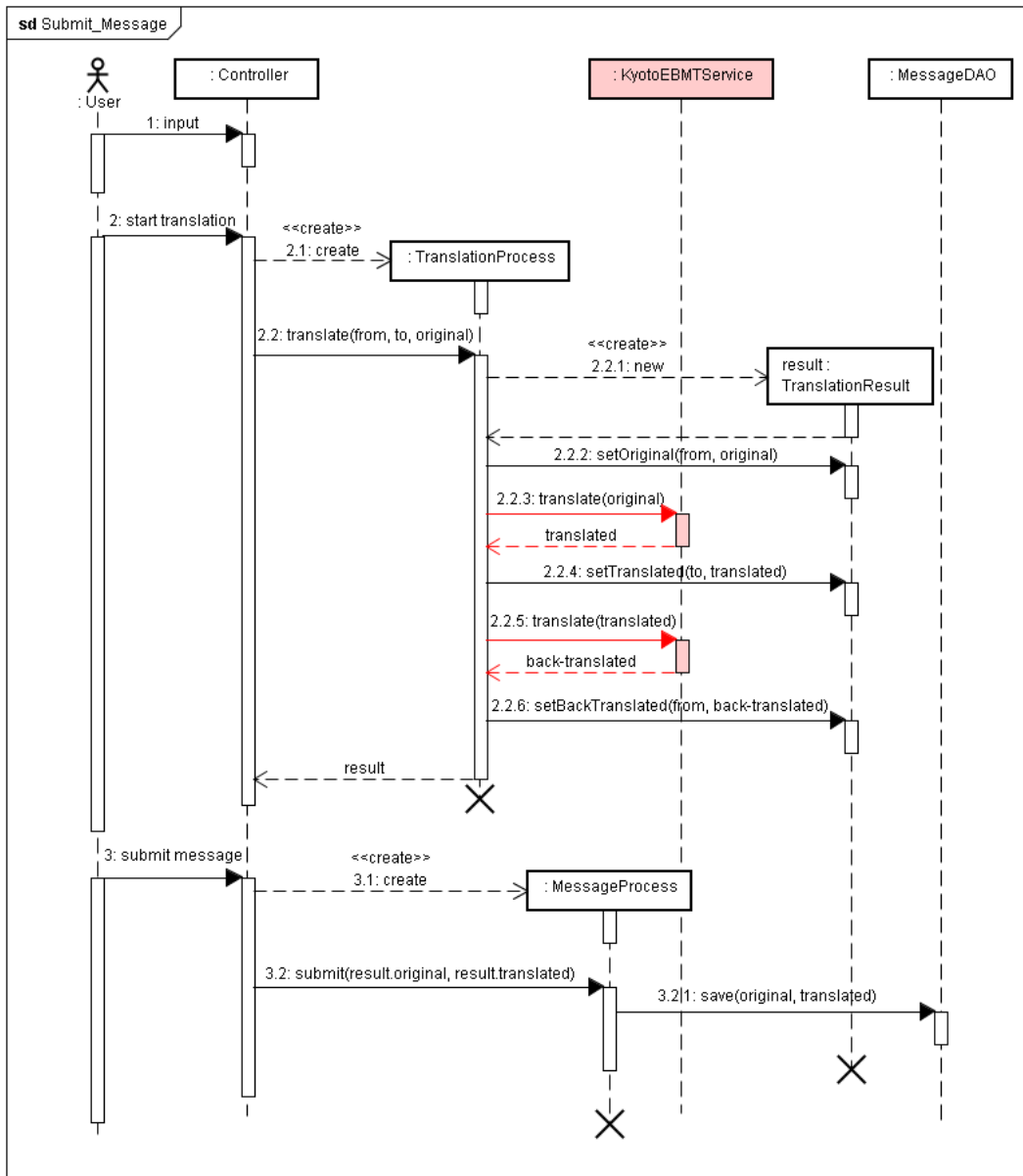


Figure 13: Sequence Diagram when Posting Message

4. When it is posted the translated message is also sent to this system and the original message and translated ones are saved as a set to the database on the system.

Receivers of the message see original or translated one decided by the language they are using. That is to say if they read in the same language with the original message, they see the original one, on the other hand if they read in a different

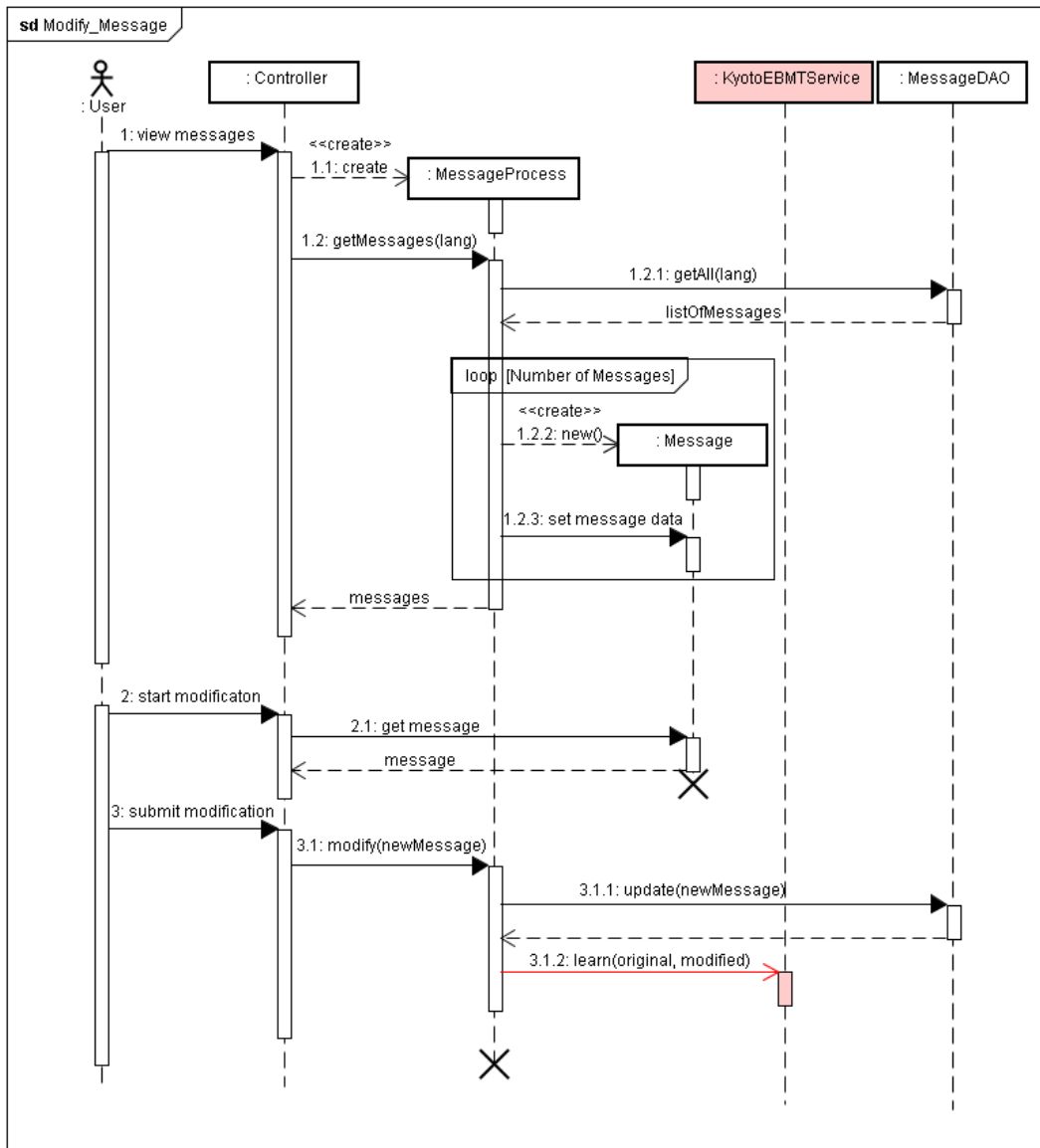


Figure 14: Sequence Diagram when Modifying Translated Sentence

language they see the translated one.

When a user modifies a message this system will update its database and let the KyotoEBMT learn the original message and the modified one as a parallel text in the flow of Figure 14. The detailed flow of this action is the following and modifications of translated messages result in the enrichment of a community corpus and improvement of the translation by the KyotoEBMT:

1. A native or near native speaker of the language of a translated message

modifies the translated message.

2. This system updates the translated message to the modified one into its database. The modified message is equivalent to the translated one so other users can modify the modified one again and the modified message will become gradually better automatically.
3. This system adds a pair of the original message and the modified one to the bilingual corpus of the KyotoEBMT and lets the KyotoEBMT learn it.

Regarding the application of the KyotoEBMT to a multilingual community I constructed the system, on the system the community corpus will be enriched automatically as described in this section. It is expected for this system to accumulate parallel texts automatically and improve the quality of the KyotoEBMT by using it in a multilingual community.

5.3 Usage of Proposed System in Multilingual Community

The advantage of the proposed system is that many usages of it in a multilingual community result in the enrichment of the community corpus and the improvement of the quality of the KyotoEBMT. The Example-Based Machine Translation, including the KyotoEBMT, needs a community corpus but it costs a lot and is difficult to prepare it in advance. On the other hand members in a community have motivations of sharing information with each other, conveying their meanings to the other member and understanding what other members want to say even if the communication in the community is conducted in multilingual. Such motivations are made clear from the result of the interview about the modification of a translated sentence to a member of NPO Pangaea as the following:

1. Why a member would like to modify a translated sentence?
 - (a) Most members in NPO Pangaea want to know the activities not only in Japan but also in Korea and Malaysia.
 - (b) After the activity held on several locations in different countries at the same time members in a certain location want to know the situation of the other locations.

- (c) Members of a certain location want to report the activity held in their location to members of the other locations.
2. What does a member do when he/she does not understand the meaning of a translated message?
 - (a) If the message should be shared really, translation volunteers are asked to translate it.
 - (b) Those translation volunteers have a motivation that they really want to help the community.

As the above there is a strong motivation to share information in a multilingual community so that each available function such as the one of the support of self-initiated repair, including the Back Translation, and the one of the modification of a translated sentence are used in communications. Regarding the modification of a translated sentence by the translation volunteers, they have motivations to really want to help the community as much as they can. So they can modify not all but some messages. The modification of a translated sentence is often voluntarily done in a multilingual community, and therefore, the proposed system realizes the support to a multilingual community by the Example-Based Machine Translation by using it well.

5.4 Application of Proposed System to Multilingual Community

With regard to the application of the Example-Based Machine Translation to an actual multilingual community I apply it to the community site that is one of practices of using the Toolbox described at 3.3 and established as one activity of the G30¹⁾ held in Kyoto University.

5.4.1 Purpose of Application

The purpose of the application of the Example-Based Machine Translation to the G30 Community Site is to verify how much it is actually useful in a multilingual community.

The Example-Based Machine Translation itself has been evaluated many

¹⁾ The “Project for Establishing Core Universities for Internationalization” (Global 30), which is established by Japan Society for the Promotion of Science.

times such as at 4.4 but has not been evaluated when it is used in an actual multilingual community yet. Or the effect of its usage in an actual multilingual community is still unclear even if it has been clear that high quality of the translation was observed in an evaluation or the translation quality was improved by the enrichment of the community corpus.

This time in this chapter I want to observe how smooth multilingual communications become by using the Example-Based Machine Translation on the proposed system in the multilingual community.

5.4.2 Application Method

The KyotoEBMT as the Example-Based Machine Translation is applied to the G30 Community Site. And large-scaled general corpora and a small-scaled community corpus are prepared in advance in a similar way to 4.3.

I use several corpora at once as a general corpora including a bilingual corpus about dialogues in tourism, which is called tourism dialogue corpus and includes about 40 thousands parallel texts, a bilingual corpus extracted from the Yomiuri Newspaper, which is called Yomiuri Newspaper corpus and includes about 250 thousands parallel texts, and a bilingual corpus extracted from bilingual dictionaries and phrase books, which is called bilingual dictionary corpus and includes about 80 thousands parallel texts. The reasons why those corpora are used even though the papers abstract corpus was used in a primary evaluation at 4.4 are the followings:

1. After improving the KyotoEBMT it was verified that those three corpora are suitable to the KyotoEBMT as learned data
2. Those three corpora, especially the tourism dialogue corpus, have more spoken expressions than the papers abstract corpus.

The following is the result of the light evaluation when those above three corpora are used at the same time as a general corpus.

The hospital reception corpus is used in the evaluation as a community corpus in a similar way to 4.4. Figure 15 shows the comparison of scores of the BLEU that is done by using two kinds of corpora as a general corpus; the papers abstract corpus, including about one million parallel texts, and the tourism dialogue corpus, the Yomiuri Newspaper corpus and the bilingual dictionary

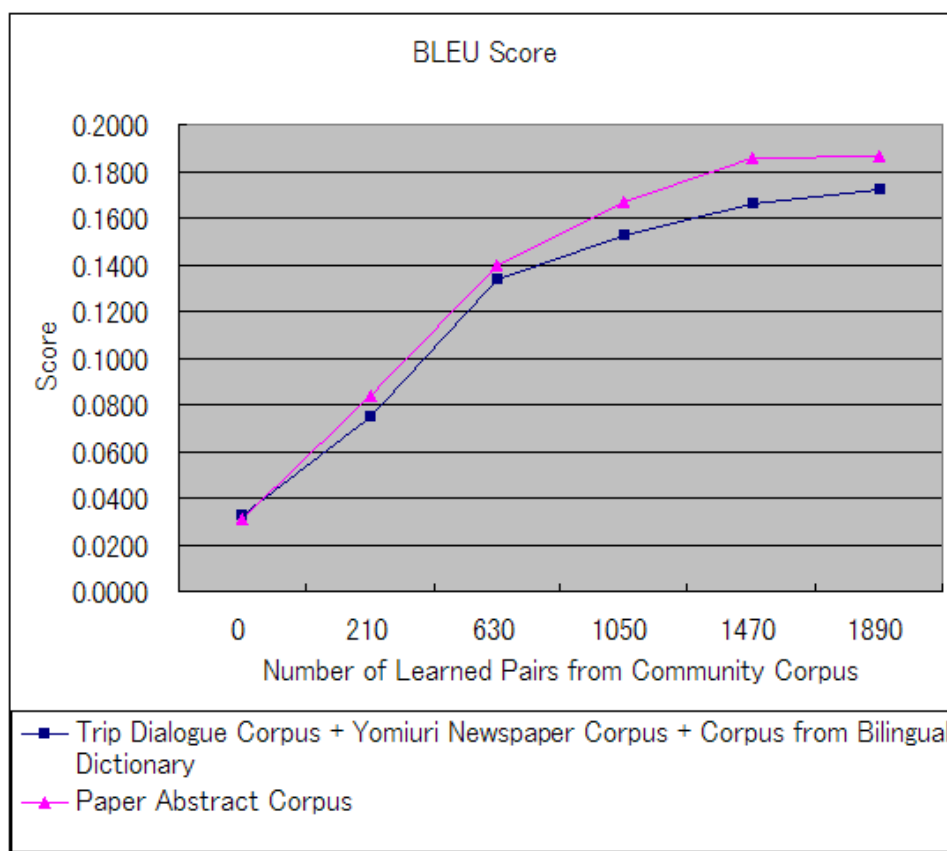


Figure 15: Automatic Evaluation by KyotoEBMT using Tourism Dialogue Corpus, Yomiuri Newspaper Corpus and Bilingual Dictionary Corpus

corpus, which include totally about 37 thousands parallel texts. This figure tells that the two have few differences with scores of the BLEU. But the speeds of the translation are very different. When the tourism dialogue corpus the Yomiuri Newspaper corpus and the bilingual dictionary corpus are used as a general corpus it takes about 4.2 seconds per sentence for the KyotoEBMT to translate. On the other hand when the papers abstract corpus it takes about 9.3 seconds per sentence. So it is appropriate that those three corpora are used as a general corpus because the quality of the translation has few differences and the speed differs about two times.

I decided to use a bilingual corpus, called Kyoto University handbook corpus, as a community corpus for the G30 Community Site. The Kyoto University handbook corpus was created by extracting parallel texts from handbooks for

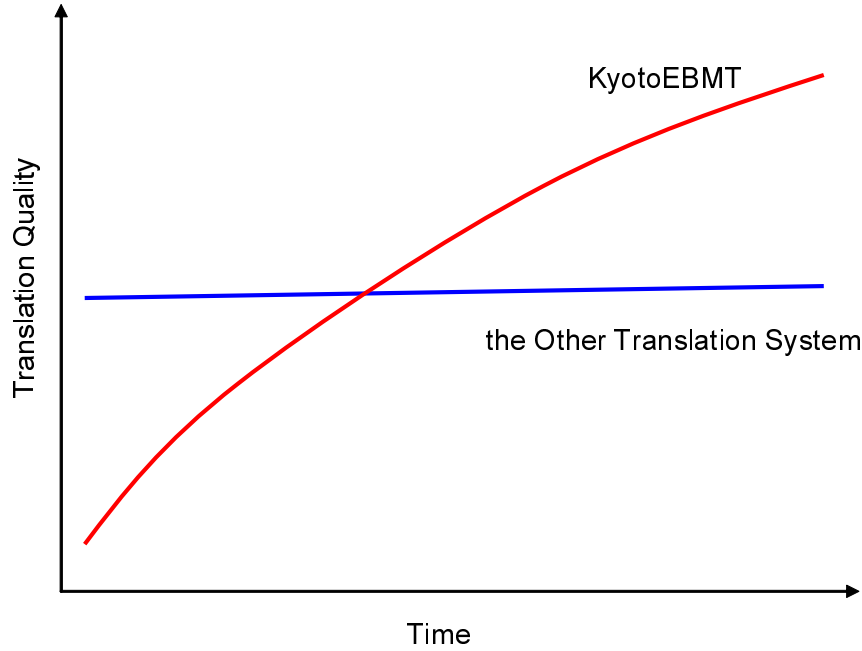


Figure 16: Transit of Translation Quality

foreign students and papers of the student affairs office. And it has about seven hundreds parallel texts in Japanese and English.

5.4.3 Devisal on Application

There is one problem about the quality of the translation when applying the KyotoEBMT to the G30 Community Site. It is that the quality or the quantity of the prepared general corpus decides the quality of the KyotoEBMT when it translates non-specific sentences to the community even if it learned a large-scaled general corpus and a small-scaled community corpus by similar way as described above. Those general sentences are often translated better by the Rule-based Machine Translation or the other translation systems. In contrast, the KyotoEBMT, however, can be customized to a community and be improved by the enrichment of the community corpus so the quality of it is expected to be raised by the consecutive use of the proposed system. Figure 16 shows the image of the transit of the translation quality. It is more desirable that the quality of machine translation is high if it is actually used in a community so I suppose that another machine translation, including the Rule-based Machine Translation, is used in combination if the KyotoEBMT cannot translate sentences well.

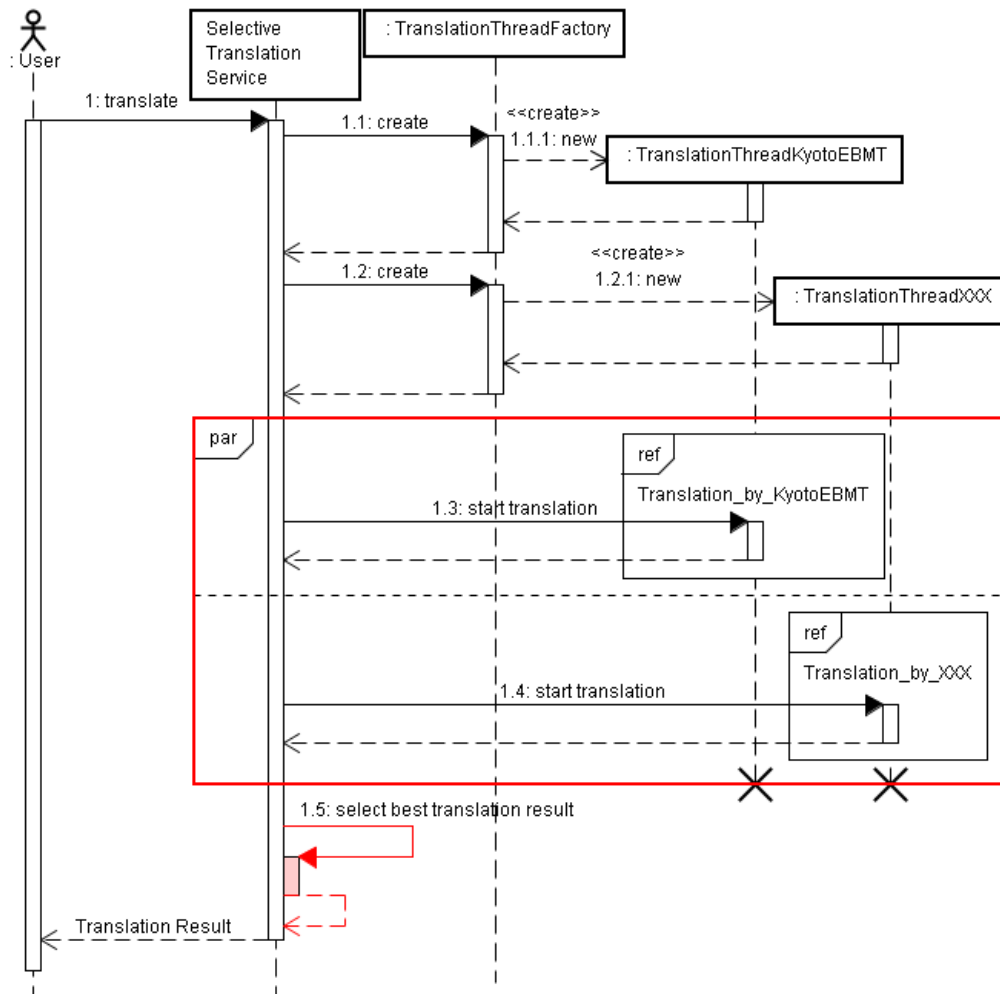


Figure 17: Sequence Diagram of Selective Translation Service

In the combination use of the KyotoEBMT and the other machine translation I constructed a selective translation service that translates sentences by using several translation services, including the KyotoEBMTService, and selects a best translation result as a response of it from results of those services by comparing results of their Back Translation to the original sentence. The flow of the selective translation service is shown in Figure 17, Figure 18 and Figure 19. In these figures the part of XXX such as the TranslationThreadXXX and the TranslationClientXXX is practically the name of the other translation system except the KyotoEBMT. And I describe the client that translates sentences by invoking the XXX service as the TranslationClientXXX in these figures. But

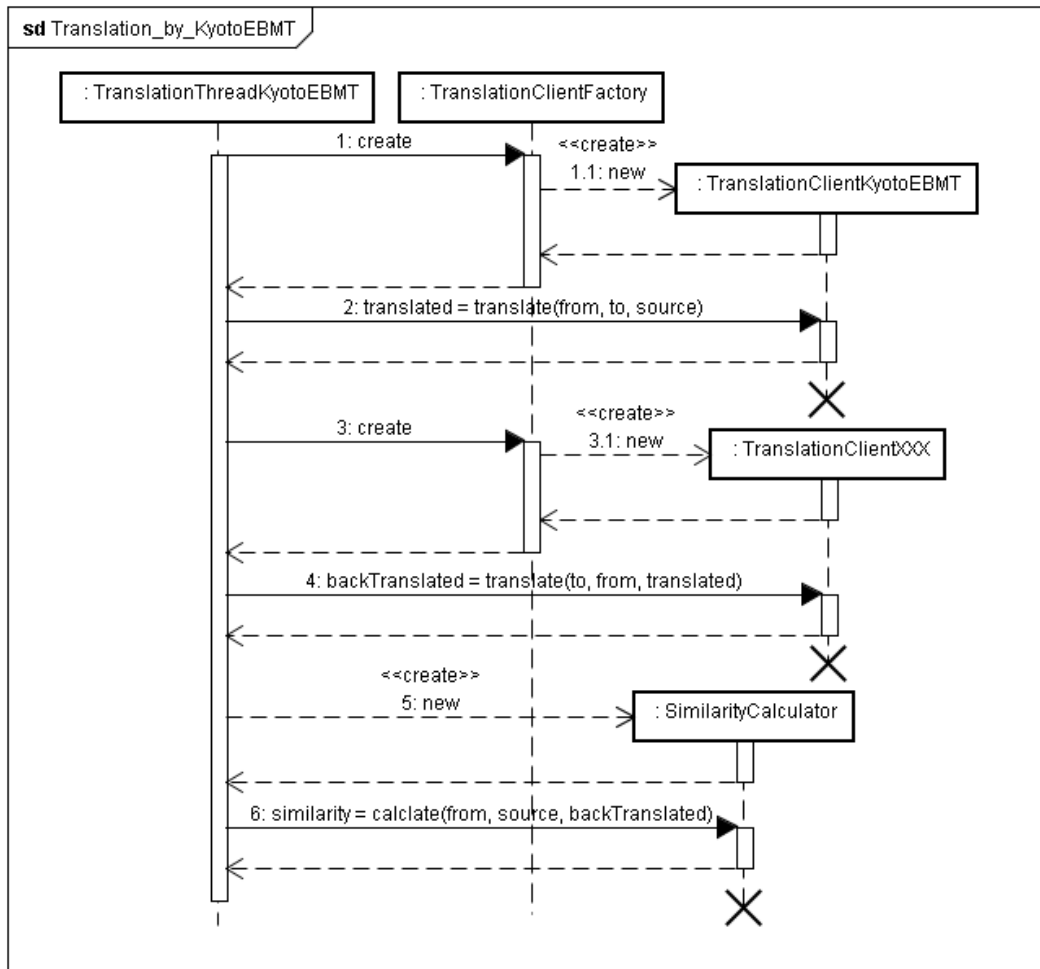


Figure 18: Sequence Diagram of Translation by KyotoEBMT

the part of actual invocation of such a service is abbreviated.

When the selective translation service receives a translation request from a user the service translates the original message by using the KyotoEBMT. At the same time the service also translates the original message by using the other translation system in the part surrounded by a red square. After translating every thread make the Back Translation of the translation result and calculate the similarity between its result and the original message. The same translation system is used in all of the Back Translation processes. And while calculating the similarity, the BLEU used in 4.4.2 is applied. So the service selects a translation result where the Back Translation has a highest score of the BLEU calculated with the original message as an answer and returns it as its response.

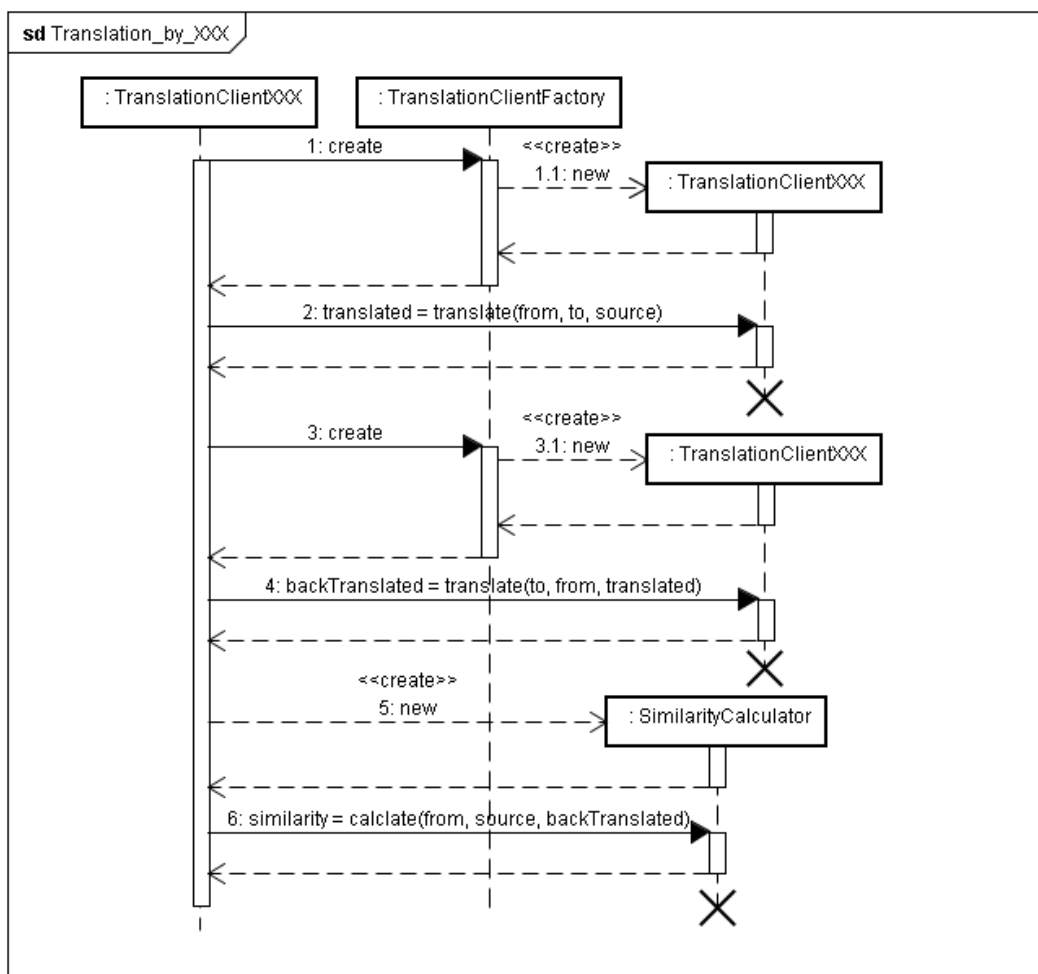


Figure 19: Sequence Diagram of Translation by the Other Translation System

5.4.4 Result of Application

The G30 Community Site was just opened in January 12, 2010, when the KyotoEBMT was applied to real use. But some messages are already posted on the BBS in the G30 Community Site and modifications of translated sentences are made. The enrichment of the bilingual corpus of the G30 Community is being observed by those modifications and it is expected that the quality of the KyotoEBMT customized to the G30 Community will be improved gradually.

Chapter 6 Conclusion

This paper focused to support an actual multilingual community by using a machine translation considering multilingual communities which opportunities are increasing in recent years as a result of globalization and the growth of the Internet. First, I introduced the Intercultural Collaboration Experiment, in which multilingual communications by using a machine translation as it was were observed, and contrived language services as the result of the experiment such as the Back Translation and the Translation with Dictionary. And I also introduced the Language Grid, which provides the framework in which those language services are easily realized, and various multilingual community support systems which have actually supported multilingual communications by using the Language Grid.

As a result of those activities, multilingual communications has become smooth by helping for humans to use a machine Translation well. For example, the Translation with Dictionary, which is the translation using a bilingual dictionary of a community while translating, makes the multilingual communication smoother.

But in actual fields of multilingual community not only specific terms to the community, which the Translation with Dictionary can translate well, but also expressions peculiar to the community, which it cannot translate well, are often used. And a problem about those expressions has left in the multilingual community.

To solve the problem I supposed that I applied the Example-Based Machine Translation to a multilingual community, which is one of kinds of machine that can be customized by preparing a bilingual corpus adequately and has a feature that the translation quality of a sentence which domain is similar to the bilingual corpus becomes high even if it is in a small scale. But there was the problem when applying the Example-Based Machine Translation that to achieve high quality of the translation by it in a community requires a large-scaled bilingual corpus of the community but it is often difficult to prepare it because of the cost to create it. So I applied the Transfer Learning to this problem about a

bilingual corpus. That is I supposed the method to create an Example-Based Machine Translation optimized to a community at low cost by preparing two types of bilingual corpora, a large-scaled general corpus which is easy to prepare and a small-scaled community corpus.

The above suggestion was enables to create an Example-Based Machine Translation customized to a community, but it is also a fact that to prepare the community corpus in a larger and larger scale makes the quality of the translation higher and higher. So I built a mechanism that more and more use of a customized Example-Based Machine Translation in a community enables to enrich the number of parallel texts in the community corpus automatically. That is to say, more and more use of it makes the quality of translation in the community higher and higher spontaneously. The mechanism is realized on the modification of the translated sentences naturally made by members of a multilingual community because of the motivation that they want to share various information. And it is expected that this mechanism achieves the optimization of the Example-Based Machine Translation to the community and high quality of the translation by using it as a result of enrichment of the community corpus.

At last, I constructed system based on the mechanism and deployed it to a community site of G30 to apply the Example-Based Machine Translation to a multilingual community actually. When the system was used it was difficult to communicate by only using a customized Example-Based Machine Translation at the early stage, so I constructed the selective translation service, which translates sentences by using several machine translations including the Example-Based Machine Translation and selects the best translation from results of those translations as a result of its response and applied it to the site. It is expected from actual observations that the modification of translated sentences enables to enrich the community corpus and the quality of the translation by the Example-Based Machine Translation used on the community site.

Acknowledgments

The author would like to express sincere gratitude to the supervisor, Professor Toru Ishida at Kyoto University, for his valuable advice and giving me the opportunity and the environment of conducting this research. The author would like to express his thanks to Associate Professor Shigeo Matsubara and Assistant Professor Hiromitsu Hattori at Kyoto University, for his continuous and valuable advice. The author would like to express his appreciations to the advisers, Professor Yuichi Nakamura, Assistant Professor Nobuaki Arai and Professor Katsuya Yamori at Kyoto University and Assistant Professor Satoshi Oyama at Hokkaido University, for valuable advice sparing their precious time. The author would like to thank all members of Ishida & Matsubara Laboratory at Kyoto University.

This research is supported by Global COE Program “Informatics Education and Research Center for Knowledge-Circulating Society” in Kyoto University. And the author is deeply grateful to members of the NICT Language Grid Project and the Language Grid Association who cooperate with him writing this paper.

References

- [1] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet(SAINT-06)* (2006).
- [2] Kamada, T. and Tanabe, N.: Possibility of Multilingual NOTA for reducing language barrier on International Collaborative Learning -The case study from International Collaborative Learning between Japan and Syria-, *International Conference for Media in Education 2008 (ICoME2008)* (2008).
- [3] Kurohashi, S., Nakazawa, T., Alexis, K. and Kawahara, D.: Example-based Machine Translation Pursuing Fully Structural NLP, *Proceedings of International Workshop on Spoken Language Translation (IWSLT'05)*, pp. 207–212 (2005).
- [4] Morimoto, S., Sakai, S., Gotou, M., Cho, H., Ishida, T. and Murakami, Y.: Building Blocks: Layered Components Approach for Accumulating High-Demand Web Services, *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 430–433 (2009).
- [5] Morita, D. and Ishida, T.: Collaborative Translation by Monolinguals with Machine Translators, *International Conference on Intelligent User Interfaces (IUI-09), Poster Session*.
- [6] Nagao, M.: A framework of a mechanical translation between Japanese and English by analogy principle, *Proc. of the international NATO symposium on Artificial and human intelligence*, New York, NY, USA, Elsevier North-Holland, Inc., pp. 173–180 (1984).
- [7] Nomura, S., Ishida, T., Yamashita, N., Yasuoka, M. and Funakoshi, K.: Open Source Software Development with Your Mother Language: Intercultural Collaboration Experiment 2002, *International Conference on Human-Computer Interaction (HCI-03)*, Vol. 4, pp. 1163–1167 (2003).
- [8] Papazoglou, M. and Georgakopoulos, D.: Service-oriented computing, *Communications of the ACM*, Vol. 46, No. 10, pp. 25–28 (2003).
- [9] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method

- for automatic evaluation of machine translation, *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, Association for Computational Linguistics, pp. 311–318 (2002).
- [10] Sakai, S., Gotou, M., Morimoto, S., Morita, D., Tanaka, M., Ishida, T. and Murakami, Y.: Language Grid Playground: Light Weight Building Blocks for Intercultural Collaboration, *International Workshop on Intercultural Collaboration (IWIC-09), Poster Session*, pp. 297–300 (2009).
- [11] Sumita, E., Iida, H. and Kohyama, H.: Translating with examples: A new approach to machine translation, *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, pp. 203–212 (1990).