# **Utilizing Agglutinative Features in Japanese-Uighur Machine Translation**

# Muhtar MAHSUT, Yasuhiro OGAWA, Kazue SUGINO, Yasuyoshi INAGAKI

Graduate School of Engineering, Nagoya University Furo-cho, Chikusa-ku, Nagoya

Japan

{muhtar, yasuhiro}@inagaki.nuie.nagoya-u.ac.jp

### Abstract

Japanese and Uighur languages are *agglutinative languages* and they have many syntactical and morphological similarities. And roughly speaking, we can translate Japanese into Uighur sequentially by replacing Japanese words with corresponding Uighur ones after morphological analysis. However, we should translate agglutinated suffixes carefully to make correct translation, because they play important roles on both languages. In this paper, we pay attention to them and propose a Japanese-Uighur machine translation utilizing the agglutinative features of both languages. To deal with the agglutinative features, we use the derivational grammar, which makes the similarities clearer between both languages. This makes our system proposed here simple and systematical. We have implemented the machine translation system and evaluated how effectively our system works.

#### Keywords

agglutinative languages, machine translation, derivational grammar, Japanese, Uighur

## **1** Introduction

Japanese and Uighur have many syntactical similarities, especially with respect to the word order in the sentence. This observation suggests that we can translate Japanese into Uighur in such a manner as word for word replacing after the morphological analysis of input sentences. In addition, both languages have morphological similarities, which is a significant feature of agglutinative languages.

However, there is a grammatical problem, that is, Uighur grammar says that Uighur has no conjugations, while traditional Japanese grammar, which is often called school grammar, says that Japanese verbs do conjugate. Therefore the word-for-word translation based on school grammar needs to deal with verbal conjugations of Japanese, in which the translation of verbal endings is very complicated to describe.

On the other hand, it is said that agglutinative language does not conjugate, and that verbal variants are derived from invariable stems by mechanically appending affixes to them. These two views contradict to each other.

To resolve this problem, G. N. Kiyose proposed the derivational grammar (Kiyose, 1995) based on the phonological approach originated by Bloch (1946). This grammar claims that Japanese verbs do not conjugate and Japanese verbal variants are made up by appending suffixes to verbal stems. Because verbal conjugations need not to be considered, the derivational grammar makes morphological analysis quite simple.

Since the derivational grammar pays attention to agglutinative features of languages, it is applicable to not only Japanese but also other agglutinative languages such as Uighur. So applying the derivational grammar to the two languages, we can make not only syntactical but also morphological similarities between them clearer. This approach also makes the translation simpler and more systematic.

Recently, there are many literatures that discuss translation between agglutinative languages such as Japanese and Korean (Kim et al., 1998; Lee et al., 1990). They use only syntactical similarities between them, but they can not make use of morphological similarities. Our approach systematizes both syntactical and morphological similarities between Japanese and Uighur through using the derivational grammar.

In this paper, we propose a Japanese-Uighur machine translation system based on the derivational grammar. First, we explain the similarities between them and show some illustrative examples of word-for-word translation. As a matter of course, the fact that Japanese and Uighur are different languages does not allow us to apply the word-for-word translation in a straightforward way. In fact, a straightforward word-for-word Japanese-Uighur translation sometimes yields unnatural Uighur sentences, which is caused by wrong combination of Uighur morphemes. We introduce an idea of morpheme replacement to solve this problem. When unnatural Uighur morpheme sequence occurs, we replace an inappropriate Uighur morpheme with an appropriate one determined by the information concerning morphemes appearing before and after it. This leads us to succeed in generating more natural Uighur sentences.

Thus we have implemented the translation system by using a Japanese morphological analyser MAJO (Ogawa et al., 1998) which we have developed by the derivational grammar approach. MAJO originally keeps the dictionary information as the list of 3-tuples <Japanese morpheme, part-of-speech, meaning>. For translation, we replace the 3-tuples with <Japanese morpheme, part-of-speech, Uighur morpheme >. As a result of it, outputs of the morphological analysis by using MAJO become word-forword translation from Japanese to Uighur. Then we apply morpheme replacement to Uighur morphemes of MAJO's outputs and acquire a translated Uighur sentence.

Our approach is not limited to Japanese-Uighur translation. We can apply our approach to translations between other agglutinative languages successively, since the derivational grammar approach treats to agglutinative features of languages.

This paper is organized as follows: Agglutinative similarities between Japanese and Uighur with examples of translations between them are shown in Section 2. Section 3 points out some problems of word-for-word translation. In Section 4, we introduce morpheme replacement rules to solve them. We describe our implementation of Japanese-Uighur translation system



Figure 1: Word-for-word translation from Japanese to Uighur

based on the derivational grammar in Section 5. Finally, we evaluate our system in Section 6.

# 2 Agglutinative Similarities between Japanese and Uighur

# **2.1 Syntactical Similarities**

Both Japanese and Uighur languages are agglutinative and they often refer to such the languages as free word order languages. In Japanese, for example, we can say "karega tobirawo aketa" as well as "tobirawo karega aketa". Both sentences mean "he opened the door" in English. What does allow such the word order changes keeping the equivalent meaning? The answer should be found in the function of case suffices. The dependency relation of a noun to other words, that is the role which plays in a sentence, is indicated by case suffixes. The Japanese case suffix '-ga' indicates the subjective noun of a sentence. That is,  $\overline{\cdot}$ -ga' makes the nominative case. The case suffix '-wo' makes the accusative case and it indicates that the prepositive noun is the object of the verb in a sentence. So case suffixes make it possible to understand sentences even if the position of "karega" and "tobirawo" are exchanged.

Uighur language has the same property. Uighur accusative case marker is '-ni' and it corresponds to Japanese case suffix '-wo'. But in Uighur language the nominative case is often indicated by the zero-form. We show it by ' $\emptyset$ '. This does not mean that Uighur language has no nominative case suffixes.

Similar by to Japanese, Uighur case suffixes make it possible to exchange word order with no change of meaning. Thus, we can translate both Japanese sentences "karega tobirawo aketa" and "tobirawo karega aketa" into Uighur in the manner of word-for-word as shown in Figure 1. Both of "u ixikni aqdi" and "ixikni u aqdi" are natural Uighur sentences that mean "he opened the door".

This observation means that the case suffixes play the essential roles in Japanese and Uighur, and should be treated carefully in Japanese-Uighur translation.

# 2.2 Morphological Similarities

It has been considered that there is a morphological difference between Japanese and Uighur. That is, although Uighur has no verbal conjugations, the traditional Japanese grammar says that Japanese verbs do conjugate. This causes a serious problem for word-for-word translation.

We overcome this problem by using the derivational grammar, which claims that Japanese verbs do not conjugate, and that appending suffixes to verbal stems makes up Japanese verbal variants. We also apply the derivational grammar to Uighur. This makes their morphological similarities clear as well as the syntactical ones.

## 2.2.1 Union consonant and union vowel

The invariable part of a verb is called a verbal stem. For example, 'tabe-' in "tabe-ru" [eat] and 'kak-' in "kak-u" [write] are both verbal stems in Japanese. There are two sorts of verbal stems: the one ending with a vowel, e.g. 'tabe-' and the one ending with a consonant, e.g. 'kak-'. We call the former a vowel stem and the latter a consonant stem, respectively.

A verbal phrase consists of a verbal stem and some verbal suffixes. The verbal suffixes of "tabe-ru" and "kak-u" are '-ru' and '-u', respectively. The verbal suffix '-ru' follows only a vowel stem and the suffix '-u' follows only a consonant stem. The derivational grammar says those suffixes are the environmental variants of the suffix '-(r)u'. The consonant 'r' appears when it is appended to a vowel stem and disappears when it is appended to a consonant stem. A consonant of this kind is called a union consonant.

Verbal phrases "tabe-na-i" [do not eat] and "kak-ana-i" [do not write] represent negative actions. The negative verbal suffixes of them are '-na-' and '-ana-', respectively. Those suffixes are also the variants of the suffix '-(a)na-'. The vowel 'a' appears when it is appended to a consonant stem and disappears when it is appended to a vowel stem. A vowel of this kind is called a union vowel.

meaning	Japanese	Uighur		
causative	-(s)ase-	-guz-		
passive	-(r)are-	-(i)I-		
potential	-(r)e-	-(y)ala-		
polite	-(i)mas-	-		
negative	-(a)na-	-ma-		
desiderative	-(i)ta-	-gu-		

Table 1: Derivational suffixes in Japanese and Uighur

form	meaning	Japanese	Uighur
finite	non-perfective	-(r)u	-[i]du
form	perfective	-(i)ta	-di
participle	non-perfective	-(r)u	-[i]digan
form	perfective	-(i)ta	-gan
converb form	perfective copulative provisional conditional negative copulative simultaneous	-(i)te -(r)eba -(a)zu -(i)nagara	-(i)p -sa -mastin -gaq
imperative	affirmative	-e,-ro	-gin
form	negative	-(r)una	-magin

Table 2: Syntactical suffixes in Japanese and Uighur

The derivational grammar summarizes these discussions as the following two rules:

- **Connection rule 1:** When the suffix beginning with a union consonant is appended to a consonant stem, the union consonant disappears.
- **Connection rule 2:** When the suffix beginning with a union vowel is appended to a vowel stem, the union vowel disappears.

## 2.2.2 Derivational Suffixes and Syntactical Suffixes

As shown in the examples of the above section, the negative verbal suffix '-(a)na-' can be followed by other verbal suffixes, for example, '-i' in "kak-ana-i" [do not write]. These examples show that the verbal suffix appended to verbal stems derives a new stem, which is called a secondary stem. A suffix that derives a new stem is called a derivational suffix. For example, "kak-ana-" is a secondary stem and (-(a)na-) is a derivational suffix. Other verbal suffixes that do not derive new stems are called syntactical suffixes. They make various verbal forms, which are classified into 4 forms: finite, participle, converb and imperative. For example, the suffix '-(r)u' in "tabe-ru" is a syntactical suffix and it forms a finite form. Japanese verbal stems can be followed by several suffixes, but syntactical suffixes are always appended last. Note that Japanese has strict rules for order of verbal suffixes, although the word order in Japanese sentence is rather flexible.

## 2.2.3 Applying the derivational grammar to Uighur

Uighur verbal phrases have the similar features to Japanese. We can also apply the derivational grammar to Uighur language. A Uighur passive derivational suffix is '-(i)l-', where '(i)' refers a union vowel. When '-(i)l-' follows a vowel stem 'yasa-' [create], a union vowel '(i)' disappears and 'yasal-' [be created] yields.

Let's consider a Uighur verbal stem yaz-' [write]. In order to express a causative meaning, yaz-' is followed by a causative suffix '-guz-' and forms a new stem 'yazguz-' [cause to write], that is, '-guz-' is a derivational suffix. In addition, '-di' completes the verbal phrase 'yazguzdi'' by appending to the new stem. Both Japanese and Uighur have several derivational suffixes and the correspondence between them is summarized in Table 2. In addition, there are some derivational suffixes other than these in Uighur. For example, a Uighur suffix '-(i)x-' indicates cooperative meaning. If 'kutra-' [congratulate] follows '-(i)x-', it forms 'kutrax-' [congratulate each other]. In Japanese similar meaning indicated by '-(*i*)*aw*-', which is not a single suffix but is combination of a syntactical suffix '-(*i*)' and a verb '*aw*-'. However, for translation, Japanese '-(*i*)*aw*-' matches Uighur derivational suffix '-(i)x-', so we take '-(*i*)*aw*-' as a derivational suffix. Similarly, we consider some compound morphemes as derivational suffixes, such as '-(*i*)*teir-*' and '-(*i*)*tutuar-*'. Both suffixes correspond to Uighur suffix '-(i)wat-'.

Here, we should note that Uighur has another connection rule. This shows a change of a phoneme instead of disappearing. For example, when the syntactical suffix '-[i]digan', which means non-perfective participle, follows a consonant stem 'yaz-', a phoneme i' appears and "yazidigan" is formed. But if it follows a vowel stem 'yasa-', 'i' changes to 'y' and "yasaydigan" is formed. We named such a phoneme a union half vowel. It varies according to the following rule:

**Connection rule 3:** When a suffix beginning with a union half vowel is appended to a consonant stem, the union half vowel becomes 'i', otherwise it becomes 'y'.

# **3** Problems for Japanese-Uighur Translation

There are many similarities between Japanese and Uighur but there are also some linguistic differences between them. Those differences cause serious problems to the word-for-word translation. In this section, some examples illustrate the problems. Our solutions to these problems are descried in section 4.

# **3.1 Problems of verbal suffixes**

Table 2 shows the correspondences between Japanese and Uighur syntactical suffixes. As you see, Japanese participle suffixes are the same form as finite suffixes while Uighur ones are not. For example, Japanese perfective syntactical suffix '-(*i*)*ta*' is used for both finite and participle forms. In Uighur, however, they are different forms as shown in Figure 2. For this reason, to translate '-(*i*)*ta*', we have to decide which of '-di' or '-gan' is the correct translation.



Figure 2: The difference about finite form and participle form

## 3.2 Problems of case suffixes

We have already explained that for Japanese and Uighur case suffixes specify the role of noun phrases in the sentences. This fact allows the flexible word order in the sentences of these languages. But there does not exist oneto-one correspondence between case suffixes of Japanese and those of Uighur. For example, "gomiwo suteru" is translated into "éhlét<u>ni</u> tökidu", where Japanese case suffix 'wo' is translated into Uighur case suffix 'ni'. On the other hand, the translation of "hasiwo wataru" is

Japanese suffix	translated Uighur suffixes and its number					
	Ø	ning	gé			
ga	3637	77	5			3719
	97.7%	2.0%	0.3%			
	ni	Ø/ni	din	others	fault	
wo	1566	320	118	57	47	2108
	74.3%	15.2%	5.6%	2.7%	2.2%	
	gé	dé	din	Ø	fault	
ni	1183	294	81	16	35	1609
	73.5%	18.3%	5.0%	1.0%	2.2%	
	dé	bilén				
de	638	14				652
	97.8%	2.2%				
	dép	Ø/dép	bilén			
to	193	46	146			385
	50.1%	12.0%	37.9%			
<u>,</u>	din	fault				
kara	289	5				294
	98.3%	1.7%				
	gé					
he	236					236
	100%					
	din					
yori	14					14
	100%					

 Table 3: Statistical correspondence between Japanese and Uighur case suffixes

"köwrük din ötidu". Japanese case suffix 'wo' is translated into 'din'. The suffix 'wo' usually indicates the object, but sometimes it indicates the place, while Uighur case suffix 'ni' and 'din' indicates the object and the place, respectively. So we have to choose 'ni' or 'din' according to the role of 'wo' in the sentence. In a fact, the case suffix 'wo' that indicates a place often appears together with such the motional verbs like "tooru" [pass] and "wataru" [cross]. This suggests we can choose 'ni' or 'din' according to the verbs which the noun phrases with 'wo' depends on.

We have investigated the ambiguities of case suffixes by using IPA Lexicon of the Japanese Language for Computers (basic Japanese verbs) (IPAL, 1987), which contains 861 Japanese verbs and 3473 example sentences. We have translated the Japanese case suffixes that appeared in the sentences into Uighur ones and counted the number of them. Table 3 shows the result. 'Ø/ni' in the second row means that 'wo' in the sentence is translated in 'ni' but it is usually omitted in the corresponding Uighur sentences. Some of example sentences can not be translated directly and we classified such cases as *faults*.

We do not deal with Japanese case suffix 'no', which is usually translated Uighur case suffix 'ning', because 'no' is usually used in such a way "A no B" [B of A] and it is related little to verbs. We do not deal with Japanese case suffix 'ya', too, because of the same reason.

### **4** Suffix Adjustment

Now let us proceed to our discussions on realization of word-for-word translation from Japanese to Uighur. The facts we revealed so far show that the problems to be solved here is how to decide verbal or case suffix correspondences correctly. To overcome these problems, we adopt a method to assign the default Uighur suffix to each Japanese suffix and then to substitute a well fitted suffix for an unnatural one under replacement rules. Since a verbal stem and a following verbal suffix affect each other, we can choose an appropriate suffix by knowing the right and left words. On the other hand, the verb which the noun phrase with a case suffix depends on affects the suffix. So we need to decide the correct case suffix

Japanese suffix	default suffix	prepositive word	postpositive word	new suffix	new parts -of-speech
-(r)u	-ydigan	* * *	end of sentence punctuation mark sentence-final particle	-ydu -ydu -ydu	finite suffix finite suffix finite suffix
-(i)ta	-gan	* * *	end of sentence punctuation mark sentence-final particle	-di -di -di	finite suffix finite suffix finite suffix

### Table 4: Replacement table of verbal suffixes

considering the verbs depended. Thus, we propose two types of rules for suffix replacement.

### 4.1 Replacement Table of Verbal Suffixes

For verbal suffixes, we make a replacement rule shown in Table 4. A Japanese suffix in the left most columns is not used for the replacement but for the purpose of helping to understand. A default suffix in the second column is a Uighur morpheme, which is translated directly from the corresponding Japanese suffix in the first column. A prepositive word and a postpositive word represent a condition for replacement of the default suffix. If the default suffix appears between the prepositive and postpositive words, it should be replaced by a new suffix in the fifth column. The mark \*(don't care) denotes that a replacement rule does not require a prepositive or postpositive word. Since words of a sentence are tagged parts of speech by the morphological analysis, we need a part of speech of the new suffix in the right most columns.

### **4.2 Replacement Rule for Case Suffixes**

There are often several phrases between a noun phrase and a verbal phrase on which it depends. So we can not use a replacement rules similar to above and need another type of rules for case suffixes. We use dependency relation and add the case pattern data to the Uighur verb, such as  $\langle \ddot{o}t - \{wo/-din\} \rangle$ . The Uighur verb ' $\ddot{o}t$ -' has a pair of Japanese case suffix and Uighur one and it is a replacement rule for case suffixes. In translation, if a Uighur verb that has such a pair appears in the sentence, the translation system searches the noun phrase depending on the verb and having the Japanese suffix contained in the pair. Then the old Uighur suffix is replaced by the new one that is in the pair.

For each Japanese case suffix, we assume the most possible one determined by Table 3 as a default Uighur suffix. But, exceptionally, we adopt ' $d\acute{e}$ ' as a default Uighur suffix for a Japanese suffix 'ni'. The reason is the following. The Japanese case suffix 'ni' often indicates a position of space or time and it is said to represent the locative case. In such a case 'ni' should be translated Uighur suffix ' $d\acute{e}$ ' which designates the locative case. On the other hand, 'ni' also denotes the dative case and corresponds to ' $g\acute{e}$ '. Table 3 shows that the dative case

'ni' is more popular than the locative one. But the locative case 'ni' is less related to verbs than the dative one. So we translate 'ni' to 'dé' as default and it is replaced with 'gé' if the verb needs an indirect object.

In addition, we consider not 'dép' but 'bilén' as a default Uighur suffix for Japanese suffix 'to', which has two meanings. One is to change the preceding sentence to a quotative clause and it corresponds to 'dép'. Another is called the comitative case, which expresses a co-operant or accompanist, and it corresponds to 'bilén'. Traditional Japanese grammar says that both designations of 'to' are case suffixes. But the derivational grammar treats the former as a conjunctive particle and only the latter as a case suffix. So we consider it is not the problem of translation but the one of morphological analysis to distinguish two type of 'to'. We use the morphological analyzer that tags only the latter as a case suffix. So we decide 'bilén' as a default for case suffix 'to'.

### 5 Machine translation from Japanese into Uighur

We have implemented the Japanese-Uighur machine translation system. Our system consists of four modules: MAJO, the two replacement modules and the morpheme connection system. MAJO is a morphological analyser of Japanese based on the derivational grammar, and its dictionary consists of 3-tuples <Japanese morpheme, part-of-speech, meaning>. For the translation, we replaced the 3-tuples with <Japanese morpheme, part-of-speech, Uighur morpheme>. Therefore outputs of MAJO become word-for-word translations for input sentences.

Here, a Japanese input sentence is "tukutta hasiwo watatta." [crossed the bridge that one constructed.] shown in figure 3. Firstly, MAJO divides it into Japanese words and yields a sequence of equivalent Uighur words. Secondly, replacement rules of verbal suffixes are applied to those Uighur suffixes if they match the conditions on the replacement table. In the example, there are two '-tta' in the input sentence and MAJO translates both into '-gan'. At this point, last '-gan' matches the replacement rule and it is replaced with '-di'. Thirdly, case suffixes are replacement rules satisfied the condition. In the example,

Japanese case suffix	ga	WO	ni	de	to	kara	he	yori
default Uighur suffix	Ø	ni	dé	dé	bilén	din	gé	din



Figure 3: Translation system and an example of translation

the first step translates 'wo' into 'ni'. But the verb 'watar' on which the noun phrase "hasiwo" depends has a replacement rule  $\langle watar$ , consonant-verb,  $\ddot{o}t-\{wo/-din\}\rangle$ . So '-ni' is replaced with '-din'. Finally, the morpheme connection system connects Uighur morphemes according to the connection rules and generates a Uighur output sentence.

### **5** Experiments

We used 136 sentences that include 306 verbal phrases (254 different patterns) to evaluate performance of our Japanese-Uighur machine translation system. We constructed a Japanese-Uighur dictionary that had about 13,000 words including 3,800 verbs. We compared translation results between the system with/without the replacement rules for verbal suffixes. As a result, the system translated 119 verbal phrases correctly without the replacement rules, while the system with the replacement rules translated 212 verbal phrases correctly. Thus, the precisions of translation about verbal phrases improved from 38.9% to 69.3%.

We also evaluate translation of 295 case suffixes appeared the 136 sentences. The simple system translated 257 suffixes correctly, while our system translates 293 case suffixes correctly. The precisions of translation for case suffixes improved from 87.1% to 99.3%.

### **6** Conclusion

In this paper, we proposed a Japanese-Uighur translation system. Our system is based on the derivational grammar and has succeeded in systematic word-for-word translation. In addition, it can generate a natural Uighur sentence by using replacement rules.

Our system has two modules for replacements and now we are going to unify them. In addition, it does not take account of amb iguities of word meaning. For example, Japanese syntactical suffix '-(r)are-' has three meanings, which are passive, potential and honorific, but we now consider it only as a passive suffix. Therefore, we need to develop a word selection method. The replacement table of our system describe only suffixes and parts of speech. If we expand it to word meanings, we could solve ambiguities of word meaning.

We have discussed only case suffixes, but there are other suffixes in Japanese, like 'mo' and 'sika'. We know some of them correspond to Uighur suffixes. For example, Uighur suffix 'mu' has similar role of Japanese 'mo'. So we should investigate their correspondence further.

We are now making bigger size of experiments of our translation system and collecting more replacement rules. We are aiming to make our system fit for practical use.

### **Bibliographical References**

- Bloch, B. (1946). Studies in Colloquial Japanese, Part I, Inflection, Journals of the American Oriental Society 66.
- IPAL (1987). IPA Lexicon of the Japanese Language for Computers (Basic Verbs), The Information-technology Promotion Agency.
- Kim, J. and Lee, J. and Lee, G. (1998). Generation of Korean Predicates Based on Modality-Feature Ordering and Lexicalizing Table in Japanese-Korean Machine Translation, Journal of Natural Language Processing 5(2), pp.3-24.
- Kiyose, G. N. (1995). Japanese grammar --A new approach--, Kyoto university press.
- Lee, H. and Nakajima, M. and Agui, T. (1990). A Japanese-Korean Predicate Translation System Using the Semantic Correspondence between Their Auxiliary Predicative Expressions, Transactions of Information Processing Society of Japan 31(6), pp.801-809.
- Ogawa, Y. and Muhtar and Toyama, K. and Inagaki, Y. (1998). Derivational Grammar Approach to Morphological Analysis of Japanese Sentences, PRICAI'98, LNAI 1531, pp. 424-435.