Comparative Study on Japanese and Uyghur Grammars for An English-Uyghur Machine Translation System

Polat KADIR Koichi YAMADA Hiroshi KINUKAWA

Dept. of Information Systems and Multimedia Design
Graduate School of Engineering, Tokyo Denki University
2 – 2 Kanda Nishiki-cho, Chiyoda-ku, Tokyo 101-8457, JAPAN
polat@cll.im.dendai.ac.jp, {yamada, kinukawa}@im.dendai.ac.jp

Abstract

Uyghur is one of the Turkic languages in the Altaic language family. We are developing a machine translation system to translate from English into Uyghur. As there are no previous researches devoted to machine translation between English and Uyghur and being short of related works that we could use as a base for our research, we noted that by making clear the morphological and syntactic similarities and differences between Japanese and Uyghur we can make use of the approaches and methods of English-Japanese machine translation to make faster progress in our research. In order to attain this goal, we have performed a comparative study on the Japanese and Uyghur grammars. In this paper, we describe the similarities as well as differences between Japanese and Uyghur in both levels of morphology and syntax and we give a brief description of our English-Uyghur transfer method to which we are aiming at applying our comparative study on Japanese and Uyghur grammars.

1 Introduction

Uyghur is the name of one of the Turkic languages in the Altaic language family and spoken mainly by the Uyghurs in the Xinjiang Uyghur Autonomous Region of China. We are currently developing a machine translation system based on a syntactic transfer method to translate from English into Uyghur (Kadir et al. 2004). However, there is not a previous work related to English-Uyghur translation that we could use as references for our research. On the other hand, Japanese, which is considered to be a member of the Altaic language family, shares a great deal of agglutinative features with Uyghur in common and there is a significant amount of similarities both in morphology and syntax between the two languages. Moreover, researches about English-Japanese machine translation have a long tradition with many achievements (Nagao et al. 1986). Under this situation, as a first step of our research, we have carried out a comparative study between Japanese and Uyghur grammars. The very purpose of this comparative study is to make clear the similarities and differences between the two languages and based on the results of our comparative study to apply the approaches and methods that have been developed for English-Japanese machine translation to our English-Uyghur translation system.

In this paper we give a presentation on our comparative study on Japanese and Uyghur, and we make a brief description of our transfer approach used in our translation system.

2 Grammatical Comparison of Japanese and Uyghur

Uyghur, like all the other Turkic languages, has a word order of *subject+object+verb* (SOV), and is considered to be an agglutinative language with very productive inflectional and derivational suffixation process in which a sequence of inflectional and derivational morphemes get affixed to a word stem. In Uyghur, a verb could have hundreds of word forms by sequentially adding different affixes to the word stem. Japanese, which is also considered to be an agglutinative language, also has the same word order and morphological features as Uyghur. Some researches show that this morphological and syntactic closeness is sufficient to obtain a relatively good translation result from Japanese into Uyghur on a transfer approach (Ogawa et al. 1997; Mahsut et al. 2001; Ogawa et al. 2000). In the following sections, we will make a comparison

between Japanese and Uyghur in two different levels: morphology and syntax with a close attention focused on their differences.

2.1 Morphological Comparison

As we compare the word formation, we could find that in both Japanese and Uyghur, word forms are generated by attaching many suffixes denoting case, mood, person, tense, etc. to one word stem as seen in Example(1).

(1) yazilmaghanliqtin ("*as it was not written*")
yaz + il +ma + ghan + liqtin
(書か+れ+ な + かった+ ので)
yaz / 書か(write): *stem*
+il / れ: passive voice
+ma / な: negation
+ghan / かった: past tense
+liqtin /ので: causal form

Generally, Japanese and Uyghur share a significant amount of morphological and syntactic features in common. However, there are also some differences in word formation of nouns, verbs, etc. In the following sections we will take a look at some aspects of word forming where Japanese and Uyghur differs.

2.1.1 Nouns

In Uyghur, when expressing "ownership", a noun is always accompanied by some grammatical categories as person, number, etc., and with different suffixes attached, a noun will express different ownership of the object (that a noun refers). Furthermore, this very same suffix will, at the same time, show different person and number categories (Tomur and Lee 2003). Table 1 shows the word "book" with two different category of person and number.

Person \ Number	Singular	Plural
1st Person	kitab-im (my book)	kitab-imiz (our book)
2nd Person	kitab-ing (your book)	kitab-inglar (your book)
3rd Person	kitab-i (his/her book)	kitab-i (their book)

Table 1: Person and Number of a Noun in Uyghur

Different from Uyghur, a noun in a Japanese sentence does not require any grammatical category or change for a word form.

2.1.2 Verbs

According to most of the Japanese grammars, a Japanese verb makes "katsuyo" (changing word forms of the verb stem) before they conjugate to show different tenses and moods, etc. But in Uyghur, there is not such inflection of verbs before conjugation. However, there are still many similarities in word form generation of verbs and most of the verbal suffixes in Japanese map the corresponding ones in Uyghur. Table 2 shows the similarities of a verb conjugation in Japanese and Uyghur (with an example of the verb "write" in two languages).

Japanese			Uyghur	
Verb Stem	Katsuyo-gobi	Suffix	Verb Stem	Suffix
書	く	ϕ	yaz	ϕ
書	か	せる	yaz	ghuz
書	か	される	yaz	ghuzul

Table 2: Verb Conjugation in Japanese and Uyghur

Still there are some differences regarding the grammatical categories of person, number and tense, etc., between Japanese and Uyghur. As in Uyghur, the concepts of singular and plural are expressed by means of the word forms of nouns (Tomur and Lee 2003), and a verb would also get different inflectional forms according to the number and person of the subject in a sentence. And at the same time, the suffixes of the verbs express number and tense. This is not the case in Japanese however as Japanese nouns do not require suffixes to express person, number, etc., and thus, there is no need for a noun-verb agreement (also see Section 2.2.4). Table 3 shows the person category of the verb in Uyghur.

	Number	Present Tense	Past Tense
1st Person	Singular	yaz-imen	yaz-dim
	Plural	yaz-imiz	yaz-dukh
2nd Person	Singular	yaz-isen	yaz-ding
	Plural	yaz-isiler	yaz-dinglar
3rd Person	Singular	yaz-idu	yaz-di
	Plural	yaz-idu	yaz-di

Table 3: The Person Category of Verb in Uyghur

When we compare the order of the affixes that are attached to a verb stem in specific order in

Verb Stem	Causitive	Passive	Aspect	Negation	Tense	Person	Modal	Mood
書く	せる	(ら)れる	ている	ない	た	φ	だろう	ね
yaz	ghuz	il	iwat	ma	idu		ghandur	he

Table 4: The Grammatical Categories of a Verb in Japanese and Uyghur

(2) 花子は 新しい 先生 について 太郎 に 詳しく 聞いて みた。
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
Hanako yengi muellim toghrisida **Taro din** tepsili sorap bakhti.

(3) 花子は 太郎 に 新しい 先生 について 詳しく 聞いて みた。
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
Hanako **Taro din** yengi muellim toghrisida tepsili sorap bakhti.

(*translation*: Hanako asked Taro about the new teacher in detail)

Japanese and Uyghur, we will find many similarities except for the following two points:

i. In Japanese, verb stem requires "katsuyo-gobi" and Uyghur verbs do not;
ii. Japanese verb forms are not dependent on the person and number of the subject in a sentence, and Uyghur verbs have different word forms according to different person, number and tense of the subject.

Table 4 shows the order of suffixes in two languages.

2.2 Syntactical Comparison

2.2.1 Word Order

Both Japanese and Uyghur can be considered as *subject* + *object* + *verb* (SOV) language, in which constituents can change order very freely as the grammatical roles of the constituents can be identified by the explicit morphological case markings on them without relying on their order.

Therefore, when we change the word order of a Japanese sentence, the word order of its Uyghur translation can be changed in same order and without a change in meaning (see Example (2), (3)).

2.2.2 The Case Category of Noun

Both in Japanese and Uyghur, case categories are expressed by means of case forms which are made by adding nominal case suffixes to nouns. The case forms in both languages show a correspondence in certain level and there is always a case particle in Japanese for an equivalent suffix in Uyghur. In Table 5, we compare the case category in Japanese and Uyghur.

Case Name	Case Particle in Japanese	Case Suffix in Uyghur
Nominative Case	は / が	φ
Possessive Case	の	-ning
Dative Case	へ	-gha
Objective Case	を	-ni
Locative Case	に	-da
Ablative Case	から	-din

Table 5: Case Categories in Japanese and Uyghur

2.2.3 The Dependency Structure of Sentences

In Japanese, the dependency structure of a sentence is usually represented by the relationship between phrasal units called "bunsetsu" and it is said that Japanese dependencies have the following rules (Watanabe Yasuyoshi et al. 2000; Kiyotaka Uchimoto et al. 1999):

i. Dependencies are directed from left to right.
ii. Dependencies do not cross.

iii. A bunsetsu depends only on one bunsetsu.

Observing the dependency structure of a sentence in Uyghur, we can also find the following characteristics that are very similar to the Japanese dependency rules above:

i. Dependency relation of a word to another is always from left to right.
ii. Dependency links between the words of a sentence do not cross.
iii. The dependent word could link to only one head word.

Because of this similarity, a word order in Japanese can be mapped to the word order in Uyghur no matter how they change.

2.2.4 Subject-Verb Agreement in Uyghur

As we have stated earlier, there is a big difference between Japanese and Uyghur in expressing the grammatical category of person and number of a noun, and in verb forms which require some affixes to express different tense of an action in a sentence. Thus, in Uyghur, to meet the subject-verb agreement in a sentence the verb puts on different inflectional forms according to the person and number of the subject, and the time of the action (see example (4), (5)).

(4) Men ete Tokyogha qaytimen.
私は 明日 東京へ 帰ります。
(*I will return to Tokyo tomorrow*)

(5) Biz ete Tokyogha qaytimiz.
私達は 明日 東京へ 帰ります。
(*We will return to Tokyo tomorrow*)

As we have stated in the previous sections, in Japanese, nouns do not require inflectional verb forms to show different person or number and thus there is no need for a subject-verb agreement in a sentence.

3 The English-Uyghur Machine Translation System

As there is not a language corpus available in Uyghur, we can not conduct our research on the base of an example based translation. Thus, we are pursuing an English-Uyghur machine translation system based on a syntactic transfer method. In this section, we briefly describe our English-Uyghur machine translation system by presenting an outline of the transfer system and a description of the transfer method.

3.1 Overview of the Translation Process

Just like most of the machine translation systems based on transfer approach, the overall translation process of our transfer system is consisted of three phases: English analysis, transformation and generation as shown in Figure 1. In our translation process, we attempt to apply the existing English-Japanese translation methods in our machine translation system so as to make our research processes faster and obtain better results. Therefore, in transfer phase and generation phase we apply some rules for Japanese to our transfer system based on our grammatical comparisons on the morphological and syntactic structures between Japanese and Uyghur from which to determine our transfer approach to English-Uyghur machine translation system. And we expect in our future work to be able to apply our translation system to other languages which has similar grammatical features as Uyghur and Japanese.

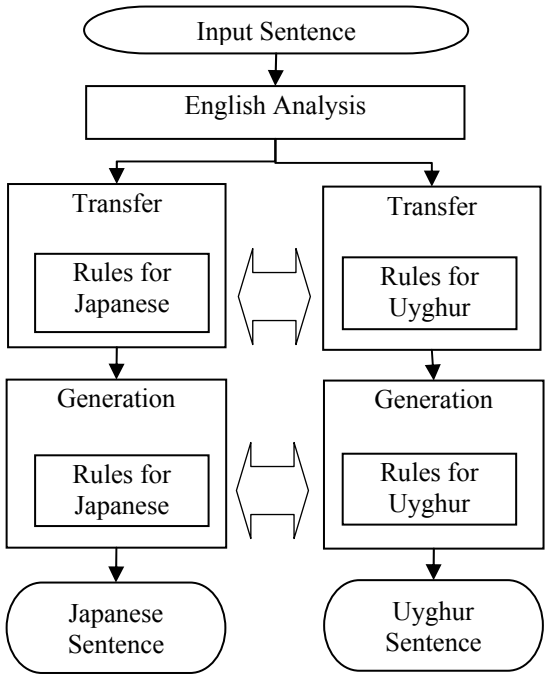

Figure 1: Overview of the Translation Process

In the following sub-sections, we will introduce our translation process by demonstrating an translation example of a simple English sentence into Uyghur.

3.2 English Analysis

In the analysis phase, we make use of Link Parser to parse English sentences to generate dependency trees of the words in the sentence. Link Parser is based on Link Grammar (Sleator and Temperley, 1991) which is proposed by

Daniel Sleator and Davy Temperley from the School of Computer Science at Carnegie Mellon University. Link Grammar is a highly lexicalized grammar with all the grammar rules defined from the words. A link grammar consists of a set of words, called the terminal symbols of the grammar, each of which has a linking requirement. The linking requirements of each word are contained in a dictionary. The Link Parser output is represented as a dependency structure of a sentence and the parsing output can be represented as a set of possible links between two words from the sentence. Table 6 shows the parsing output of the English sentence "*I bought a book at a store.*"

I	← Sp →	*buy*
bought	← MVp →	*at*
bought	← Os →	*book*
a	← Ds →	*book*
at	← J →	*store*
a	← Ds →	*store*

Table 6: The Link Parser Output

In the table above, the word pairs on the left and right ends of every row are linked by a link type indicating a dependency relation of the word pairs identified by uppercase letters with some of which followed by lowercase letters.

3.3 English – Uyghur Transfer

In the transfer phase, all the words from the parsing result which is expressed as a set of links are replaced by corresponding Uyghur words so that we can get a word replacement result as in Table 7.

men	← Sp →	*setiwal*
setiwal	← MVp →	*da*
setiwal	← Os →	*kitab*
bir	← Ds →	*kitab*
-da	← J →	*dukkan*
bir	← Ds →	*dukkan*

Table 7: Word Replacement

Due to the very different morphological and syntactic structure that English and Uyghur have, most of the occasions some changes of the word order are needed between verbs and their objects and between affixes and nouns, and also some words will have to be deleted or added as in our cases in Table 7. This is done by referencing the linking requirements of each word which are specified in Uyghur dictionary.

After the word order manipulation, as could be seen from the word replacement results in Table 7, the phrase structure representation in the table also needs to be transferred into a proper order from which we can generate a syntactic structure of Uyghur. In order to achieve this, we manipulate the orders of phrases of the table such that the same word on the left most end of a row comes on the right most end of the next lower row as shown in Table 8.

men	← S →	*setiwal-*
-da	← MV→	*setiwal-*
dukkan	← J →	*-da*
bir	← D →	*dukkan*
kitab	← O →	*setiwal-*
bir	← D →	*kitab*

Table 8: Syntactic Transfer

Based on this linkset transformation results, we can build up a translated sentence-like structure with a proper word order as in Example (6).

(6) Men bir dukkan-da bir kitab setiwal-

3.4 Uyghur Sentence Generation

The grammar rules that we use for generating Uyghur sentences are defined in exactly the same way as the Link Grammar rules of English words by specifying the linking requirements of each word in the lexicon. And the linking requirements for each word are expressed as a formula involving the connector names, operators "*&*", "*or*", and parentheses. In Example (7), (8), we describe some simple word rules for some nouns and verbs that are used in a test for generating Uyghur sentences:

(7) moshuk sut kitab oy:
D- & (S- or O+ or J-) or {@A-};

(8) yaz setiwal yugur:
S- & {J+} & O+ & W-;

Generating an output sentence of Uyghur from the structure in Example (6) is done by referencing the linking requirements of the words defined throughout the word dictionary of Uyghur.

One of the biggest differences between Uyghur and English is that Uyghur has a word order of *subject+object+verb* and is an agglutinative language with very productive inflectional and derivational suffixation process. In Uyghur, a verb

could have hundreds of word form by adding different suffixes to the verb stem as demonstrated in Example (1) in section 2.1. In the generation phase, an Uyghur sentence is generated by operations such as generating suffixes and determining tense, mood, person, number and case particles etc. of a verb of the sentence.

In the case of our sentence structure produced in the transfer phase, we need to add certain suffixes to supply the proper grammatical category of person and number and case suffixes of nouns, and tense, mood and persons in verb forms which are required to produce a final sentence that makes sense in Uyghur as in Example (9).

(9) Men dukkan-da bir kitab setiwaldim.
(*I store at one book bought*)

4 Conclusion

The motivating idea behind this contrastive study is to emphasize the difference of Japanese and Uyghur in order to make better use of the achievements of English-Japanese translation in our research. In this paper, we proposed to utilize the approaches and methods of English-Japanese machine translation research results in our English-Uyghur machine translation system. To implement this approach in our translation process we have conducted comparative study on Japanese and Uyghur grammars. When the comparison of meaning is considered, we will need to make more specific study in semantic mapping between Japanese and Uyghur. In our future work, we will need more detailed observation of this cross linguistic differences.

English and Uyghur belong to distinct language families with diversity of differences regarding the word formation rules and syntactic rules. English, considered to be a weakly inflected language, depends mostly on word order in determining the semantics. While in Uyghur, inflectional and derivational suffixes determine the semantic change on the root word in a regular and predictable way. When generating a translation output sentence of Uyghur from the transfer phase, choosing appropriate suffixes to attach to the verb stem and determining tense, mood, person, number and case particles etc. makes up the most complicated part of the translation process.

This is a report about the work in progress on English-Uyghur machine translation system and there is still much work to do. In our future work, we plan to make a further study of the rules for transferring and generating Uyghur sentences based on the comparative work we have done, develop the English-Uyghur dictionary, implement the English-Uyghur machine translation system and make evaluations on the system in the near future.

References

Polat Kadir, Koichi Yamada and Hiroshi Kinukawa. 2004. *An English-Uyghur Machine Translation System*. In "Proceedings of The 66th National Convention of IPSJ", pages 51-52, Information Processing Society of Japan, Tokyo, Japan

Makoto Nagao, Jun-ichi Tsujii, Jun-ichi Nakamura. 1986. *Science and Technology Agency's Mu Machine Translation Project*. In "Future Generations Computer Systems 2", pages 125-139, North-Holland, Amsterdam, Netherlands

Yasuhiro Ogawa, Muhtar Mahsut, Katsuhiko Toyama and Yasuyoshi Inagaki. 1997. Japanese-Uighur Machine Translation based on Derivational Grammar: A Translation of Verbal Suffixes, *IPSJ SIG-Notes,* NL-120-1

Yasuhiro Ogawa, Muhtar Mahsut, Kazue Sugino, Katsuhiko Toyama and Yasuyoshi Inagaki. 2000. Verbal Phrase Generation based on Derivational Grammar in Japanese-Uighur Machine Translation, *Journal of Natural Language Processing*, 7(3): 57-77

Muhtar Mahsut, Yasuhiro Ogawa and Yasuyoshi Inagaki. 2001. Translation of Case Suffixes on Japanese-Uighur Machine Translation, *Journal of Natural Language Processing*, 8(3):123-142

Hamit Tomur and Anne Lee. 2003. *Modern Uyghur Grammar*. Yildiz, Istanbul, Turkiye

Yoshiyuki Watanabe, Shigeki Matsubara, Katsuhiko Toyama, Yasuyoshi Inagaki. 2000.Einichi Douji Tsuuyaku-no Tame-no Zenshinteki Nihongo Seisei, *Proceedings of The Sixth Annual Meeting of The Association for Natural Language Processing* , pages 272-275

Kiyotaka Uchimoto, Satoshi Sekine, Hitoshi Isahara. 1999. *Japanese Dependency Structure Analysis Based on Maximum Entropy Models*. In "Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics", Bergen, Norway, pages 196 – 203

Daniel Sleator and Davy Temperley. 1991. *Parsing English with a Link Grammar*. In "Carnegie Mellon University Computer Science technical report CMU-CS-91-196