The Development of Japanese-Uighur Machine Translation System

MAIMITILI NIMAITI　　IZUMI YAMAMOTO
Department of Computer Science and Engineering
Graduate School of Engineering Nagoya Institue of Technology
ciq17639@stn.nitech.ac.jp izumi@nitech.ac.jp

1 Abstract

In this paper,we discuss about the different type of machine translation system and we propose a simple and systematical translation system. We are developing a machine translation system to translate from Japanese into Uyghur. As there are no previous researches devoted to Hybrid machine translation system between Japanese 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 the old rule base Japanese-Uighur machine translation and other opensource Statictical machine translation system 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

2 Introduction

Uyghur is the name of one of the Turkic languages in the Altaic language family and spoken mainly by 8.5 million (2004) Uyghurs in the Xinjiang Uyghur Autonomous Region of China. Uyghur is also spoken by 300,000 in Kazakhstan, and there are Uyghur-speaking communities in Afghanistan, Albania, Australia, Belgium, Canada, Germany, Indonesia, Kyrgyzstan, Mongolia, Pakistan, Saudi Arabia, Sweden, Taiwan, Tajikistan, Turkey, United Kingdom, USA, and Uzbekistan. There are two main languages in Xinjiang Uyghur Autonomous Region: Uyghur and Mandarin Chinese. Mandarin Chinese is not used widely in southern Xinjiang. About 80 newspapers and magazines are available in Uyghur; 22 TV channels and ten publishers serve as the Uyghur media. The necessity of machine translation system is growing rapidly due to the increase of government documents and translation request.Since the Japanese and Uighur languages are agglutinative languages and have many syntactic similarities. For the first generation of the translation system which translate from Japanese into Uighur languages by replacing Japanese words with corresponding Uighur words after Japanese morphological analysis.

We are currently developing a machine translation system based on a syntactic transfer method to translate from Japanese into Uyghur (Kadir et al. 2004). However, there is not a previous work related to Hybrid machine translation system between Japanese and Uyghur 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. In this paper we give a presentation on our comparative study on Japanese and Uyghur.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

3 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.

3.1 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)kuralmiganliktin("as it was not seen")

kur + al +mi+ghan + liqtin

(見ら れ な かった ので)

kur/見ら（see）:stem

+al/れ:passive voice

+mi/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.

3.1.1 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 shows the word " time " with two different category of person and number.

	Singular	Pulural
1st Person	wakit-im	wakit-imiz
2nd Person	wakit-ing	wakt-inglar
3rd Person	wakit-i	wakt-i

3.1.2 verb

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.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. When we compare the order of the affixes that are attached to a verb stem in specific order in Japanese and Uyghur, we will find many similarities except for the following two points:

i. In Japanese, verb stem requires " katsuyogobi " and Uyghur verbs do not;

ii. Japanese verb forms are not dependent on the person and number of the subject in a sentence, and

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

Uyghur verbs have different word forms according to different person, number and tense of the subject.

3.2 Syntactical Comparison

3.2.1 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.

3.2.2 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.

3.2.3 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.

3.3 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 subjectverb 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 (2), (3)).
(2)man hazir tamak yayman.
私は 今 ご飯 食べます。
(3)biz hazir tamak yaymiz.
私達は 今 ご飯 食べます。
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.

4 Hybrid Machine Translation

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT companies (Asia Online and Systran) are claiming to have a hybrid approach using both rules and statistics. The approaches differ in a number of ways.Rules post-processed by statistics: Translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine.Statistics guided by rules: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to postprocess the statistical output to perform functions such as

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

normalization. This approach has a lot more power, flexibility and control when translating.

5 Future work

As a new member of Machine Translation family. Japanese-Uighur Machine Translation system has been developed by few ways. There are also a few works and resource in the past which can rely on it.But until now all the systems are based on Rule based Machine translation system.Those project also is less portable.In future. We propose to collect the resourse and develop the Hybrid Machine Translation system that can use those related resource and try to make it portable system which can run in any system as Windows,Mac Os and Linux.

6 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 Hybrid machine translation system between Japanese and Uyghur in our research . 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 cross linguistic differences. 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 Japanese-Uyghur dictionary, implement the Japanese-Uyghur machine translation system and make evaluations on the system in the near future.

7 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

"Machine Translation An Introductory Guide", Arnold, Balkan, Humphreys, Meijer, Sadler, 1994.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.