PARTICLE-BASED MACHINE TRANSLATION FOR ALTAIC LANGUAGES : THE JAPANESE-UIGHUR CASE

Muhtar MAHSUT†, Fabio CASABLANCA†, Katsuhiko TOYAMA‡ and Yasuyoshi INAGAKI†

† Department of Information Engineering, Nagoya University, Furo-cho,Chikusa-ku
Nagoya 464-01, JAPAN
‡ School of Computer and Cognitive Sciences, Chukyo University Tokodate 101, Kaizu-cho
Toyota 470-03, JAPAN

ABSTRACT

In comparison to the many researches carried out on Machine Translation between European languages, or between Japanese and English, not many studies have focused on Japanese and other languages. In this paper we show that, when Machine Translation is performed between Japanese and a language which belong to the same family, i.e. the Altaic family, we can benefit from common characteristics.

Taking Japanese and Uighur as a study case, we investigate in detail the correspondence relation between particles, which are a shared feature of the two languages. We show that it is possible, by means of the particle correspondence relation, to establish a correspondence between the syntactic structure of the sentences and to deal with ambiguous sentences. The obtained results are expected to play an important role in Japanese-Uighur Machine Translation systems.

INTRODUCTION

Machine Translation (MT from now on) has been a very challenging field in the area of Natural Language Processing for many years. The very early approaches were largely unsuccessful, not only for lack of computing resources, but also because the complexity of the interaction effects in natural languages phenomena had been underestimated. MT is an applied area which benefit from advances in the area of theoretical Artificial Intelligence and Natural Language Processing - in spite of some partial results which have been achieved, we are still far from a satisfying treatment of natural language.

So far, in Western countries, the major part of the results have been obtained on MT between European languages, which share a similar structure; in Japan, the researches have focused on MT between Japanese and English. Examples of recent classes of MT systems which are being studied are Example-Based MT[1], Knowledge-Based MT[2], Lexical-Based MT[3, 4], Statistics-Based MT[5, 6], and so on. Of course, some of the characteristics of these models depend on the investigated languages.

At the moment, not many studies have focused on MT inside the family of Altaic languages, to which Japanese belong. First of all, all these languages are strongly *agglutinative*, i.e. form new words by combining separate parts which have their own meaning. More importantly, bound forms, like particles and auxiliary verbs, (in Japanese, *fuzokugo*) are very developed and play an important role in the sentence. We think that the *fuzokugo* structure is an important clue that deserves attention in MT between Altaic languages.

For this purpose, we propose in this paper a new approach to MT , which we call Particle-Based MT. As a case study, we examine the Japanese and Uighur* language, which both fall in the class of Altaic languages[7, 8] and present similar structures in the syntax and the lexicon.

We investigate the correspondence relation between particles in the two languages and show that it is the base of syntactic correspondence between the two languages, and hence of a MT system which shows a high degree of precision.

In the next section we discuss the functions of the particles in Japanese and Uighur and in Section 3 we

*The Uighur language, which belongs to the subfamily of the Turkic languages, is spoken, by a 1986 estimation, by 6.750.000 people. 6.500.000 speakers live in Xinjang, Western China, 245.000 in former Soviet Republics and few thousands in Afghanistan and Mongolia. The Uighur people is one of the five main nationalities in China.

describe in detail the correspondence relation between particles in these two languages. In Section 4 we discuss how ambiguous sentences can be solved paying attention to the particles. Finally, in Section 5 we conclude with a discussion of our approach.

FUNCTION OF THE PARTICLES IN JAPANESE AND UIGHUR

In agglutinative languages, such as the family of Altaic languages, fuzokugo plays a crucial grammatical role. In fact, the particles, although do not have any lexical meaning by themselves, are to decide the syntactic role of the conceptual words in a sentence. In particular, when a particle is added to a word, it adds a new meaning to the word, or it shows the relation of the word with the others in the sentence.

In Japanese and Uighur, not only the particles are important component parts, but they have very similar functions.

For example, given a sequence of the Japanese conceptual words **taro, jiro, hanako, shoukaisuru**. Here not only the sequence does not have a meaning, but we have no clue on the relation between the words. However, adding the case particles **ga, ni, wo**, we obtain a sentence which has a complete meaning:

JS* : taro **ga** jiro **ni** hanako **wo** shoukaisuru.
ES: Taro will introduce Jiro to Hanako

Taking note that *shoukaisuru*, to introduce, translates in the Uighur word *tonuxturidu*,and that to the Japanese particles **ga, ni, wo** correspond the Uighur ones **ø,gä,ni**, in Uighur we have:

US: *taro* **ø** *jiro* **gä** *hanako* **ni** *tonuxturidu.*

We can see that a word becomes subject, object depending on its particle; also, particles are used to specify the direction of the action.

On the other side, the order of the words in a sentence of the languages as Japanese or Uighur is rather free; for example, we can change the order of the words in the two sentences above:

JS′ : taro **ga** hanako **wo** jiro **ni** shoukaisuru.
US′ : *taro* **ø** *hanako* **ni** *jiro* **gä** *tonuxturidu.*
JS″: hanako **wo** jiro **ni** taro **ga** shoukaisuru.
US″: *hanako* **ni** *jiro* **gä** *taro* **ø** *tonuxturidu.*
JS‴: hanako **wo** taro **ga** jiro **ni** shoukaisuru.
US‴: *hanako* **ni** *taro* **ø** *jiro* **gä** *tonuxturidu.*

Although there are some differences in nuance between the sentences , they state identical factual things. Once more, it is evident that the syntactic structure of a sentence of the languages is entirely decided by the particles.

By contrast, in languages such as English, the order of the words is very important and the syntactic structure depends upon the words' order in the sentence. Let us look at the two English statements below:

ES1: The policeman arrested the thief.
ES2: The thief arrested the policeman.

In spite that they both contain the same words, the sentences have a quite different meaning.

CORRESPONDENCE OF PARTICLES

Showing the correspondence relation between the particles amounts to establish a correspondence of the syntactic structure of Japanese and Uighur languages, which is a key factor in achieving an efficient MT.

In modern Japanese, the particles are classified into 6 classes according to their functions in the sentence[9]:

- *case particles(kakujoshi)*:**ga, no, ni, wo, he, de, to, kara, yori, ya**;
- *conjunctive particles(setsuzokujoshi)*: **node, kara, ba, ga, te, temo, shi, noni, keredo**(mo);
- *adverbial particles (fukujoshi)*: **made, dake, gurai, sae, sura, shika, zutsu, nado, bakri, hodo, ka, yara** ;
- *topic particles(kakarijoshi)*: **ha, mo, koso** ;
- *final particles (shuujoshi)* : **ka, na(naa), sa, yo, ne, tomo**
- *interjective particles (kantoujoshi)* : **ne, sa, ya**.

Generally, the *case particles* are added to the end of a noun to represent its relation to other words in the sentence. The *conjunctive particles* are added to the end of a declinable word or of an auxiliary verb to join it to the rest of the sentence. The *adverbial particles* when added to a word affect the remaining part of the sentence as an adverb. The *topic particles* are added to a noun and affect the following predicate. The *final particles* are added to the end of a sentence

*JS, US and ES show Japanese,Uighur or English sentences respectively.

to express its intensity, an interrogation, an impression, or a prohibition. The *interjective particles* are placed at the pauses of a discourse to call the attention of the hearer, and may again convey various nouances.

A finer classification of the particles is possible according to a finer observation of the their functionalities. We recognize three main possibilities.

- the same particle belong to more than one class. For example, the particles **kara** and **ga** can be either case or conjunctive particles. In this case, we may consider the case-**kara**(**ga**) and the conjunctive-**kara**(**ga**) to be different particles because they are neatly distinguishable.

- case particles can distinguished according to the cases they might be used for. For example, the case particle **de** can be used, beyond the others, for the place-case(*bashokaku*), instrument-case(*dougukaku*) or material-case(*zairyokaku*).

- case particles may be used for the same case, but are not completely interchangeable. For example, **de** and **ni** can be used for the place-case, **kara, yori,wo** can be used for the departure-case(*shuppatsukaku*).

  The following two sentences show interchangeability:

  JS1 : kare ga ie **wo** deta.
  JS2 : kare ga ie **kara** deta.
  ES1-2: He left his room. (JS1 = JS2)

  However, in the following two sentences **ni** and **de** are not interchangeable.

  JS3: gakkou no mae **ni** ginkou ga aru.
  ES: There is a bank in front of school.
  JS4*: gakkou no mae **de** ginkou ga aru.
  (JS4* is a incorrect sentence.)

A correspondence between Japanese and Uighur particles need to respect these classifications. For the Japanese language, we examined the detailed functions of the particles by reference to an exhaustive Japanese dictionary[9]; for the Uighur language, in absence of a similar dictionary , we used the native speaker's competence of the first author. A complete correspondence function between particles has been obtained; for lack of space, Table 1 shows only the correspondence between case particles. It has been possible to obtain almost always a one-to-one correspondence. In particular, the particle **de** which is difficult to treat in MT system between Japanese and English is corresponding one-to-one to the particle **d***ä* of Uighur language. For particles which correspond to more than one particle, we notice that, while it is nice and remarkable to have one-to-one correspondence, more important is the possibility of representing the functions of a particle of a language in another one exactly. It is often quite difficult, if not impossible, to express the functions of a Japanese particle by one or two English words.

In Table 1 there are also cases where to a single Japanese particle correspond more than two Uighur particles (for example, in the case of particle **ni**, to which correspond the two Uighur particles **da,ga**).

In a MT system, we may have to decide which one to choose. We believe that this problem can be solved. Firstly, for some Japanese particles it is possible to know in which situations they match a given Uighur particle. Secondly, as to any Japanese and Uighur verbs is associated a case pattern, corresponding particles may be chosen according to the case pattern.

Especially the case particles of the two languages, which are involved with the surface and depth structures, are very important. Table 2 summarizes the particles according to the cases. While in Uighur to a case corresponds exactly one particle, the same is not true for Japanese.

AMBIGUITY OF DEPENDING PARTICLES

Ambiguity is perhaps the greatest problem in Natural Language Processing, especially in MT[10]. We may have ambiguity both at the syntactic and at the semantic level; in order to resolve it , several methods have been proposed[11] , such as the context-based or the knowledge-base techniques. However, there are particularly "hard" ambiguous sentences which can not be solved, if not looking at the causes of the ambiguities. We will show that our approach may be useful in the case of syntactic ambiguity.

Beyond ambiguity on the conceptual words in the two languages, there are two main sources of ambiguity between Japanese and Uighur:

1. **Ambiguity of the surface case**: the surface case is an important information to formalize the syntactic structure of Japanese and Uighur languages. Although there are various interpretations, we have to choose only one of them, the most appropriate, and it is not always so straightforward. Let us consider the following sentence:

Japanese particles	functions	Uighur particles	Japanese illustrative sentences	Uighur illustrative sentences
de	instrument	**d***ä*	gakkou he densha **de** iku	mäktäp kätiralway **d***ä* barmaḳ
	material	**d***ä*	tsukue wo ki **de** tsukuru	üstälni yaǧaq **d***ä* yasamaḳ
	location	**d***ä*	toshokan **de** shiryou wo shiraberu	kütüphana **d***ä* matiriyal izdämäk
	time	**d***ä*	3 jikan **de** ikeru	3 saät **d***ä* baralamaḳ
	cause	**d***ä*	kouzui **de** ie ga taoreta	säl **d***ä* öy örüldi
	manner	**d***ä*	sugoi hayasa **de** tobasu	ajayip sür'ät **d***ä* uqarmaḳ
	means	**ø**	3 nin **de** iku; jibun **de** yaru	3 adäm **ø** barmaḳ; özi **ø** ḳilmaḳ
no	modifies a noun to show possession, source, partition	**ning**	kawa **no** nagare; watashi **no** ie	därya **ning** eḳimi; mi **ning** öyüvm
	shows the agent or the object of a participal adjective	**ø**	kare **no** kaita tegami; mizu **no** nomitai hito	u **ø** yazǧan hät; su **ø** iqküsi bar adäm
	makes of a participal adjective a noun	ax,äx, ix, *...	iku **no** wo yameru	bar**ix** ni boldi ḳilmaḳ
	shows explanation or conclusion	none	saigo no chansu datta **no** da	äng ahirḳi pursät idi
	shows possession	**ning**	kore ha boku **no** da	bu mi **ning**
	shows similarity	**ø**	yama **no** youda	taǧ**ø** däk ḳilidu
ga	shows the subject	**ø**	kawa **ga** nagareru; koko **ga** gakkouda	därya **ø** aḳidu; bu yär **ø** mäktäp
	shows the object of expressions of feelings or possibilities	**ø**	mizu **ga** nomitai hana **ga** kireida eigo **ga** dekiru	su **ø** iqküsi ba r güllär **ø** qiraylıḳ engiliz tili **ø** bülidu
ni	location	**d***ä*	niwa **ni** aru; higashi no hou **ni** aru	ḥoyli **da** bar; xärḳtäräp **t***ä* bar
	moment	**d***ä*	7 ji **ni** okiru	7 **d***ä* orundin turmaḳ
	goal	**gä**	toukyou **ni** tsuku	tokyou **gä** yitip barmaḳ
	purpose of the action	**gä**	hon wo kai **ni** iku	kitap alix **kä** barmaḳ
	basis	**gä**	sono genri **ni** yotte handan dekiru	xu pirinsip**kä** asasän ḥökümd k ilǧli bolidu
	partner	**gä**	tomodachi **ni** au	dosti **gä** yoluḳmaḳ
	result of change	**ø**	gakusha **ni** naru	elim-pän hadimi **ø** bolmaḳ
	shows association, combination	**gä**	enpitsu **ni** nooto **ni** pensaki wo kudasai	ḳäläm **gä** hatir **gä** ḳäläm uqini ḳoxup bering
	manner	**gä**	aomuke **ni** taoreru	ongdisi **gä** yiḳilmaḳ
	rate	**gä**	sannin **ni** hitori no kyousouritu	üq adäm **gä** bir adämning talixix nisbeti
	cause	**d***ä*	shiken **ni** kurushimu	imtiḥan **d***ä* muxäḳḳät qäkmäk
	shows the comparison basis	**din**	otouto **ni** masaru	inisi **din** üstün turmaḳ

Table 1 Particle Correspondence in Japanese and Uighur (continue)

*Noun form of the predicate words in Uighur language

he	shows a turn of direction	**g***ä*	kochira **he** nagero	buyni **g***ä* at
	shows an arrival point	**g***ä*	toukyou **he** iku	Tokyo **g***ä* barmaḳ
	shows the noun toward which the action is directed	**g***ä*	kokoroatari **he** toiawaseru	k*öğä*ldiki **g***ä* m*ä*sliḥ*ä*t salma ḳ
	shows a state	**g***ä*	dekakeyou to shita yokoro **he** okyaku ga kita	sirtḳa qiḳay dig*ä*n y*ä*r **g***ä* miḥman k*ä*ldi
wo	object	**ni**	ongaku **wo** kiku	muzika **ni** anglamaḳ
	shows a place of passage	**din**	gakkou no mae **wo** tooru	m*ä*kt*ä*p aldi **din** *ö*tm*ä*ḳ
	shows a departure point	**din**	kuukou **wo** shuppatsu suru	ayrudurum **din** yol*ğ*a qiḳmaḳ
	shows direction	**ni**	gooru **wo** mezasu	nixan **ni** ḳ*ö*zl*ä*m*ä*k
to	shows association or coordination of actions or situations	**bil***ä***n**	otouto **to** asobu	ine **bil***ä***n** oynamaḳ
	shows result or conclusion	**ø**	yakenohara **to** natta	k*ö*y*ü*p ḳaḳas **ø** boldi
	shows quotation	**ø**	”iku” **to** itta	barim*ä*n **ø** didi
	shows association or coordination of nouns	**bil***ä***n**	kami **to** enpitsu	ḳ*ä*ğ*ä*z **bil***ä***n** ḳ*ä*l*ä*m
	shows the object of a comparison	**bil***ä***n**	kore **to** kurabete minasai	buning **bil***ä***n** selixturup k*ö*r
	shows state	**ø**	sassa **to** katadukeru	gaqqid*ä*k **ø** yi*ğ*uxturmaḳ
ya	shows association, coordination	**ḥ***ä***m**	kami **ya** enpitsu	ḳ*ä*ğ*ä*z **ḥ***ä***m** ḳ*ä*l*ä*m
yori	shows a comparison basis	**din**	kore **yori** are no hou ga ii	buning **din** awu yahxi
	asserts an exception in the context of a negative sentence	**din**	sore **yori** shikata ga nai	uning **din** baxḳa qar*ä*yoḳ
	shows the departure point of an action	**din**	gogo 6 ji **yori** hajimaru	q*ü*xtin kiyin sa*ä*t alt*ä***din** baxlanidu
kara	shows a departure point	**din**	chihou **kara** joukyou suru	t*ö*w*ä*n **din** Tokyo g*ä*k*ä*lm*ä*k
	shows material	**din**	wain ha budou **kara** tsukuru	*üzü*m ḥari*ğ*ini *üzü*m **din** yasaydu
	shows source	**din**	sensei **kara** homerareru	mu*ä*llim **din** t*ä**ğ**d*irl*ä*nm*ä*k
	shows cause	**din**	koufun **kara** nakidasu	ḥayajanlinix **din** yi*ğ*law*ä*tm*ä*k
	shows passage	**din**	mado **kara** hairu	d*ä*riz*ä***din** kirm*ä*k

Table 1 Particle Correspondence in Japanese and Uighur(end)

Case	Subjective	Genitive	Objective	Dative	Departure	Instrument	Place	Time
Japanese	ga	no	wo	he,ni	kara yori,wo	de	de,ni	ni
Uighur	ø	ning	ni	*gä*	din	*dä*	*dä*	*dä*

Table 2 Correspondence of particles in Japanese and Uighur languages in relation to the main cases.

JS1: *keisanki* **de** *honyaku shita bunshou wo kouseisuru.*

There are two interpretations for the JS1:

ES1′ : (Someone), using a computer, corrects a letter which is translated by someone.
ES1″ : (Someone) corrects a letter which is produced from a computer

In appropriate situations, both sentences may be correct.

2. **Ambiguity of the parallel structure**. Although this ambiguity is present also in English, the absence of plural makes it more relevant in Japanese and Uighur. The following sentence is a good example for this kind of ambiguity.

JS2: *taro* **to** *jiro* **no** *joushi* **ga** *sensei* **to** *atta.*

There are three possible interpretations for the part in italic of JS2:

JI1 : (taro) **to** (jiro **no** joushi)
EI1 : Taro and (Jiro's boss)
JI2 : (taro **to** jiro) **no** joushi
EI2 : (Taro and Jiro)'s boss
JI3 : (taro **no** joushi) **to** (jiro **no** joushi):
EI3 : (Taro's boss) and (Jiro's boss).

In order to do the correct selection, we need to have supplementary information about the relation between Taro and Jiro. However, to make this information available may be complex, if not unfeasible.

3. **Ambiguity about the meaning structure**. This ambiguity is observed in a sentence when its syntactic structure can be solved, but not its meaning structure. For example, we have,for the phrase:

JS3 : *kare* **no** *shousetsu*

three alternative interpretations:

EI1 : the novel that he has got
EI2 : the novel that he has written
EI3 : the novel about him

In this case, even if we do lexical-semantic processing, it is not possible to disambiguate the sentence.

We find that all these ambiguities depend on the particles **de**, **to** and **no**. This suggests that we can define a correspondence relations between Japanese and Uighur particles, not only we can avoid to perform the syntactical analysis, but we can carry the ambiguity of the Japanese sentence into the Uighur sentence and viceversa, leaving the final interpretation of the sentence to the reader.

For example, the Japanese particles **de** , **to** and **no** which are the causes of ambiguity in examples JS1,JS2, JS3, can be directly translated into the Uighur particles **dä**,**ning** and **bilan**, which absolve to the same functions. At this point the obtained Uighur sentence will have the same content of ambiguity as the original Japanese sentence and there will be no need to perform any analysis on the ambiguity.

CONCLUSION

In this paper, we have investigated the correspondence between particles in two languages, Japanese and Uighur, which belong to the Altaic family, and we have shown that the correspondence relation which can be inferred is very useful for a MT system between the two languages.

We are now implementing a Japanese-Uighur MT system based on the Particle-Based MT approach here proposed. The system is starting to show promising results.

As a final observation, we would like to notice that the Japanese language is surely the most studied of Altaic languages, in the field of Natural Language Processing. The promising results obtained by our work suggest that the large amount of scientific results and

applications available for Japanese might be adapted to other languages belonging to the same family.

REFERENCES

1. E.Sumita, H. Iida and H. Kohyama,"Example-Based Approach in Machine Translation", in: *Proceedings of the 1990 IPSJ Conference*, Information Processing Society of Japan, (1990).

2. M.Nagao,"Dependency Analyzer: A Knowledge-Based Approach to Structural Disambiguation", in: *Proceedings of Thirteenth International Conference on Computational Linguistics*(1990), 282-287.

3. B.J. Dorr,"Machine translation: a view from the lexicon", *The MIT Press*(1993).

4. J.Tujii, and K.Fujita, "Lexical Transfer Based on Bilingual Signs: Towards Interaction During Transfer", in: *Proceedings of the European Chapter of the Association for Computational linguistics*(1991), 275-280.

5. P.F. Brown, "A statistical Approach to Machine Translation", *Computational Linguistics*, Vol.16, No.2(1990), 79-85.

6. S.Doi and K.Murakami, "Translation Ambiguity Resolution Based on Text Corpora of Source and Target Languages", in: *Proceedings of Fourteenth International Conference on Computational Linguistics*(1992), 525-531.

7. K.Mabuchi (ed.), "Origins of the Japanese Language:-An International Collection of Essays", *Musashino Shoin* (1986) *(in Japanese and English).*

8. T.Kamei et al.(eds.), "Encyclopedia of Linguistics", *Sanseido* (1988)*(in Japanese)* .

9. T.Umezao et al.(eds.), "The Great Japanese Dictionary", *Koudansha*(1989)*(in Japanese)* .

10. K.Nagao and H.Maruyama Hiroshi, "Ambiguities and their Resolution in Natural Language Processing", *Jouhou Shori*, Vol.33,No.7, Information Processing Society of Japan(1992), 746-755 *(in Japanese).*

11. T.Ehara and H.Tanaka, "Natural Language Processing in Machine Translation", *Jouhou Shori*, Vol.34,No.10, Information Processing Society of Japan(1993), 1266-1273*(in Japanese).*