Paraphrasing Honorifics

Kiyonori OHTAKE and Kazuhide YAMAMOTO
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288, JAPAN
{kohtake, yamamoto}@slt.atr.co.jp

Abstract

This paper reports on a paraphrasing method for Japanese honorifics. Japanese honorific expressions, as seen in real world dialogs, have many forms of identical meanings. This paper discusses a paraphrasing method that simplifies each utterance by removing honorifics. To simplify an utterance, we take a practical approach: investigate a corpus, and construct paraphrasing rules that eliminate honorifics. We discuss how the constructed paraphrasing rules are effective for the simplification of each utterance, and a disambiguation method for some honorific verbs that require disambiguation in order to be paraphrased.

1 Introduction

Honorifics are one of the characteristic properties of spoken languages. In general, it is said that language is a communication tool, but this might imply the transmission of intellectual information. In fact, language conveys various information in many ways, and it is here where honorifics play a very important role to express emotional information.

The processing of honorifics will be very important in the future, especially in the field of spoken language translation (SLT). This is because the basic methods or techniques in SLT for practical communications will be more mature, necessitating the processing of emotional information to achieve more natural communications using translators in various situations.

In this paper, we discuss a paraphrasing method for Japanese honorifics; however, it might also be applied for Korean since Korean has honorific language that bears a striking resemblance to the Japanese honorifics. For example, the Japanese expression *koko ni aru* (to be there) corresponds to the Korean *yeogi issda*, and one of its honorific forms *koko/kochira ni gozaimasu* corresponds to *yeogi isseupnida.* In addition, Javanese and Hindi also have honorifics that are similar to the Japanese honorifics. Therefore, we hope that our discussion of Japanese honorifics will contribute towards developments in processing honorifics for Korean, Javanese, or Hindi.

Japanese honorifics can be expressed by various forms that are fundamentally the same in terms of meaning, to express one's honor, respect, and other emotions. People utilize one of a number of forms of honorifics according to the degree of honor or respect, addressee, or situation.

In processing honorifics, there are two directions: from ordinary form to honorific form and vice versa. The important point to remember is that any one ordinary form and its honorific forms differ in formation, but both of them convey fundamentally the same intellectual information. Therefore, the most suitable processing of honorifics is paraphrasing.

In this paper, we discuss a paraphrasing of Japanese honorifics that simplifies each utterance. There are many honorific forms corresponding to one ordinary form. As the first step of paraphrasing honorifics, we attempt to paraphrase honorific forms into their ordinary forms. If various honorific forms can be paraphrased into their ordinary forms, the utterances become easier to process, both for programs and humans.

The simplification of honorifics is applicable to SLT in various situations. In SLT, we earlier proposed a new paradigm that emphasizes monolingual processing for both the source and target languages (Yamamoto et al., 2001). This paradigm attempts to resolve a majority of the existing translation problems by twin paraphrasing processes. The above two directions of paraphrasing honorifics can be reflected in the twin paraphrasing processes: from honorific forms to their ordinary forms and the reverse can be reflected in the paraphrasing of the source language and that of the target language, respectively.

2 Honorific Forms in Japanese

In this paper, we focus on the following five classes as honorific forms, some of which were earlier introduced by Kaiser et al. (2001).

- Regular subject-honorific forms: formulations of verbs that exalt the subject of its utterance.
- Regular humble forms: formulations of verbs that exalt the object of its utterance.
- Irregular subject-honorific and humble forms: irregular forms of subject-honorific and humble forms.
- Polite forms: polite formulations for non-verbal expressions.
- Euphemisms: forms that are used to avoid direct negative expressions or make utterances more polite.

2.1 Regular subject-honorific forms

A subject-honorific form is an expression that exalts the subject of its utterance. There are two types of regular subject-honorific forms. One is *o/go-* V *ni naru*, where V is a verb stem, and the other is a passive form. Verbs that have special (irregular) forms do not usually have regular equivalents. In the structure *o/go-* V *ni naru*, V is sandwiched between *o/go-* and *ni naru.* If V is a verbal noun, *ni naru* replaces *suru.* The choice between the honorific prefixes *o-* and *go-* basically depends on whether the item the selection will attach to is native-Japanese (*o-*) or Sino-Japanese (*go-*). However, there are some exceptions, such as *o-denwa* (telephone), which is Sino-Japanese.

Passive forms can be used as slightly less polite honorifics. Examples of regular subject-honorific forms are as follows:

Ordinary form: *anata ga ronbun wo* ***kaku*** (You **write** a paper.)

o/go- V ni naru **form:** *anata ga ronbun wo* ***o-kaki ni naru***

Passive form: *anata ga ronbun wo* ***kakareru***

There are some fixed expressions that look like *o/go-* V *ni naru*, but are in fact not. In other words, there are some exceptions of honorific forms that take the formation: *o/go-* V *ni naru*, but are not subject-honorific. For example, *watashi wa iroiro na hito ni* ***o-sewa ni naru*** (I **am looked after** by all sorts of people) has the expression *o-sewa ni naru*, but the subject of this utterance is *watashi* (I), who cannot be exalted. The subject *watashi* should never be exalted in honorific language.

2.2 Regular humble forms

A humble form is an expression that exalts the object of its utterance. The regular humble formation takes the form of *o/go-* V *suru.* The choice between *o-* and *go-* is conditioned by the same factor as mentioned above under Section 2.1. In addition, verbs that have special (irregular) forms do not usually have regular formations. Another humble formation takes the form of *o/go-* V *mousiageru.* The form, *mousiageru*, is the irregular humble form of *iu* (to say). However, this formation has the same function as the form *o/go-* V *suru.* This formation gives us a more formal impression.

2.3 Irregular honorific and humble verb forms

A number of commonly used verbs that refer to a person's action are not used in their regular honorific forms; instead, a different 'specialized' honorific verb is used. Moreover, some honorific verbs can be used for more than one action: *meshiagaru* is used for both eating and drinking, and *irassyaru* is used for coming, going, and being there. Accordingly, we have to disambiguate these honorific verbs when paraphrasing them.

Table 1 shows these irregular verbs (for slots where no irregular verbs exist; regular formations are given in parentheses).

Table 1: Major irregular subject-honorific and humble verbs

Ordinary	Honorific	Humble
iru	*irassharu*	*mairu*
'to be'	*o-ide ni naru*	
iku	*irassyaru*	*mairu*
'to go'	*o-ide ni naru*	
morau	(*o-morai ni naru*)	*itadaku*
'to receive'		*chōdai suru*

2.4 Polite forms

So far, this paper has described honorifics of verbs, whereas polite forms are non-verbal expressions. To express one's politeness, a person changes a non-verbal expressions into its polite form. For example, to refer to a person apart from title or pronoun, the person may use *kata*, the honorific equivalent of *hito* (person), in the singular and the reduplicated *katagata* (persons) in the plural. An example is *taihen kenkō na* ***kata*** *da* (He is a very healthy **person**). The forms *kata* and *katagata* are expressions without honorific prefixes.

In contrast, to express one's politeness, a person attaches the prefix *o-* or *go-* to a noun that is unlimited. In general, native speakers of Japanese do not attach these prefixes to their own property. Examples include, *o-hashi* (chopsticks), *o-aji* (the taste), etc. However, some people always use the prefix *o-* or *go-* even when referring to their own property. For example, *watashi no* ***o-saifu*** *ga nusumare mashita* (My **wallet** was stolen). Honorific nouns of this type are called "beautified" words in Japanese. The use of "beautified" words varies greatly from individual to individual.

Not only nouns but also other words have polite forms. For instance, adjectives: *yoi/ii* (good) has the form *yoroshii*, *atsui* (hot) has the form *o-atsui*, etc.; and copula expressions: *da* has the form *desu*, and so on.

2.5 Euphemisms

There are a great number of expressions that seem to be euphemisms. By avoiding direct negative expression or employing euphemistic phrases, these euphemisms, indirect expressions, or periphrases make utterances more polite.

For example, in some cases, Japanese negative ending form *nai* can be expressed by the verbal suffix *kaneru* to avoid a direct negative expression: *hoshou deki****nai*** (I can **not** guarantee it.) can be expressed by *hoshou si****kaneru***. In other cases, various euphemistic phrases can be employed to express one's politeness when expressing one's thought, ask a favor of someone, etc. Some examples of euphemistic phrases are as follows: *kore de ii* ***to omoi masu*** (**I think** this will be fine.), *mado wo akete kure* ***nai deshou ka*** (**Could you please** open the window?), and so on.

All of the expressions shown above are euphemisms. To raise the degree of politeness further, some other expressions are utilized. One of them is to make an introductory remark before requesting someone to do something, such as ***sumimasen ga,*** *mado wo akete kure masen ka* (**Excuse me,** could you open the window?).

3 Paraphrasing Japanese Spoken Language

Japanese honorifics have a huge variation in expressions. The most typical parts of Japanese expressions are predicative parts. Therefore, spoken Japanese has a huge variation in expressions compared with the written language.

We can make many honorific forms from one utterance to reflect the corresponding social relationships in the real world or emotions of the speaker. For example, from the utterance "*mado wo akero* (Open the window)," we can make "*mado wo akete kudasai*," "*mado wo akete kure masen ka*," "*mado wo akete itadake naidesyou ka*," etc. From the examples shown above, we realize that honorifics are one of the features of spoken languages, and at the same time, they cause an enormous variety of expressions.

If various utterances having basically the same meanings could be made simpler, the utterances could become easier to process, both for programs and humans. At first glance, honorific language may appear to have a rigid grammar. However, there are many exceptions or expressions that are hard to handle as honorifics due to their being idiomatic expressions.

In order to tackle this problem, we take the practical approach of attempting to simplify and make direct expressions. Before considering a paraphrasing method, we have to clarify the targets of the paraphrasing and to know what we will paraphrase. We first attempt to collect paraphrasing phenomena by replacing functional words as much as possible, by observing a spoken language corpus. In this work, ATR SLDB (Speech and Language DataBase) (Morimoto et al., 1994) is utilized as the analysis target. This collection of texts contains formal travel type conversations between two persons. The purpose of the conversations in the corpus is to acquire some information from a clerk, or to claim something to a clerk. Therefore, the dialogs include many sentences expressing the speaker's emotions, questions, and intentions.

As a result of an analysis of the corpus, we found that there are three types of expressions that should be paraphrased.

1. honorifics
2. formal-styled language
3. phonemic changes

We discuss the paraphrasing of each expression in the following.

3.1 Paraphrasing honorifics

There are many honorific forms in the corpus because the dialogs involve many kinds of question-answer conversations, and there are many honorific expressions most of which are spoken by the clerk. All of the types we mentioned in Section 2 are included in the corpus. The real cases and their paraphrases are as follows:

Regular subject-honorific forms: The expression *o-tameshi ni naru* (to try) should be paraphrased to *tamesu.* Regular subject-honorific forms also have passive-formed honorifics. However, we do not handle these types of passive-formed honorifics in this paper, since it is hard for us to determine whether the passive form is used as an honorific form or really as a passive form. In addition, passive forms sometimes express the possibility of an action, for example, *mirareru* (to be able to see, to see, or to be seen) in *yama wa* ***mirare*** *mashita ka* (**Could** you **see** the mountain?), *yama wo* ***mirare****ta no desu ka* (**Did** you **see** the mountain?), or *dare ka ni* ***mirare*** *mashita* (I **was seen** by someone.). Accordingly, we leave this problem for our future work.

Regular humble forms: The expression *o-shirabe shi masu* (I will check it) should be paraphrased to *shirabe masu.*

Irregular honorific and humble verb forms: The expression *ossyaru* (to say) should be paraphrased to *iu.*

Polite forms: The expression *o-isya san/sama* (medical doctor) should be paraphrased to *isya.*

Euphemisms: The expression *o-tomari negau koto ga deki masu* (You can get accomodations) should be paraphrased to *tomare masu.*

Honorific language seemingly has a very restricted grammar. However, there are a number of exceptions, and many forms that should not be paraphrased due to their being idiomatic expressions. In addition, Japanese speakers sometimes misapply a number of honorifics.

3.2 Paraphrasing formal-styled language

Formal-styled language does not express the speaker's politeness, honor, and respectfulness. However, from the viewpoint of reducing variations, we consider paraphrasing formal-styled language. The concrete cases are as follows: the expression *(go-)iriyou* (necessity) is a formal styled expression of *hitsuyou*, the expression *honjitsu* (today) is a formal styled expression of *kyou*, and so fourth.

3.3 Paraphrasing phonemic change

Although there are few basic patterns in this type, they are frequent in spoken Japanese. A major pattern is to change '*no*' to '*-n-*', as in *shita****no*** *desu* (I did it) to *shita****n*** *desu*, where the former expression is the normal pronunciation while the latter is informal and colloquial. Another pattern is the omission of 'i,' as in *shiteita* (I had been doing it) to *shiteta*, also a normal to colloquial change.

4 Paraphrasing Method

At the moment, and at least in the task of spoken language paraphrasing, it is not the time to seek the automatic acquisition of paraphrasing rules. The reason for this is that, unlike tagging, parsing, and other natural language processing (NLP) applications, the target and goal of paraphrasing are unclear now. Moreover, the phenomenon of paraphrasing itself in the real world is also uncertain. We plan to explore what we can do by paraphrasing first, rather than how we can construct paraphrasing rules.

Our paraphraser consists of two components: a POS-based paraphraser and a verbal-feature-based paraphraser. The reason why there are two paraphrasers is that rules for the POS-based side are easily written, but the paraphrasing is limited. Therefore, easy rules and complex rules are divided into two modules. In addition, we construct a disambiguation part due to some irregular honorific verbs that require disambiguation.

It has been observed that most of the phenomena seen only in spoken languages change locally. Considering this, we focus on the local changes of linguistic phenomena into more ordinary expressions.

4.1 POS-based paraphraser

The overview of this paraphraser is as follows:

1. segment and part-of-speech (POS) tag by JUMAN[1], and parse (or chunk) by KNP[2]
2. convert to a labeled string (see Table 3 for examples of labeled utterances)
3. attempt to apply all of the paraphrase patterns once, in an order given in advance
4. repeat Step 3. if a pattern can recursively be applied to the result of Step 3.

We first segment, tag, and parse the input utterance. We use its top hierarchy of the POS system as it is. However, some labels are separated from the top hierarchy of the POS system to meet paraphrasing requirements: verbal nouns are isolated from nouns, suffixes are ramified under verbal, nominal, and numerical classifiers, and particles are classified into four subclasses.

The results of the analysis are formatted to a sequence of morphemes and their POSs, where one POS has one assigned character as listed in Table

[1] http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

[2] http://pine.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

2 and chunking is expressed as spacing. Table 2 shows 20 POSs in all, among which 14 are originals in JUMAN's top hierarchy and six are ramified.

Table 2: Parts-of-speech and their symbols

a: adjective, b: copula, c: conjunction, d: adverb, e: sentence-final particle, g: nominal suffix, h: prefix, i: interjection, j: numerical classifier, k: demonstrative, n: noun, p: (other) particle, q: conjunctive particle, r: attribute, s: verbal noun, t: verbal suffix, u: case particle, v: verb, x: auxiliary verb, z: symbol

The current version of the paraphraser is implemented by *Perl*. Each pattern is written in *Perl*'s substitution command `s///`. In total, there are 548 patterns, each constructed by hand, where one pattern is one `s///`. Table 3 shows an example of the paraphrasing processing.

4.2 Verbal-feature-based paraphraser

The difference between the POS-based paraphraser and this verbal-feature-based paraphraser is the target of the paraphraser. The paraphrasing rules by *Perl*'s substitution command are very easily written. However, these rules cannot easily handle inflections of predicates, because to inflect predicates correctly, we need to know the inflection types and forms.

We constructed almost 30 rules by hand, and show some examples of these rules: if the utterance has a sequence: '*o*- CV *kudasai*,' then it is paraphrased to 'CV-*te kudasai*,' where CV represents a continuous-formed verb and CV-*te* represents a *te*-typed continuous-formed verb. Concrete examples, in this case, are as follows: *o matchi kudasai* (Just a moment./ Can you hold on?) is paraphrased to *matte kudasai*, *o tanoshimi kudasai* (Please, enjoy yourself.) is paraphrased to *tanoshinde kudasai*, and so on. Another rule is that if the utterance has a sequence: 'VN *shi kane masu*,' where VN stands for verbal noun, then it is paraphrased to 'VN *deki masen*.' A concrete example is as follows: *hosyou si kane masu* (I can not engage) is paraphrased to *hosyou deki masen*.

This paraphraser also segments, tags, and parses the input utterance like the POS-based paraphraser. The paraphraser reads the parsed utterance one by one, and checks all rules in an order given in advance.

4.3 Disambiguation for some irregular honorific verbs

We mentioned in Section 2.3 that some irregular honorific verbs need disambiguation. So far, many disambiguation methods have been proposed and discussed (Yarowsky, 2000). However, in the field of disambiguation for irregular honorific verbs, there are some aspects different from conventional disambiguation problems. The most different point is that the verbs that should be disambiguated are limited to dozens at most. Accordingly, we take a deterministic approach towards disambiguation as an easy solution.

The disambiguation method utilizes a dependency structure of the utterance and some clues, and then decides the sense of the verb. The current disambiguation method examines whether two chunks, that are the closest to the target verb and depends on it, include some clues or not. Table 4 shows examples of disambiguation rules. We manually constructed 18 rules for 10 verbs based on SLDB observations.

Table 4: Examples of disambiguation rules

target verb & clues	determined sense
target: *ukagau*	
conjunctive particle *ka*	*tazuneru* (to ask)
tsuite	*kiku* (to listen)
sochira ni	*iku* (to go))
target: *irassyaru*	
dochira/doko kara	*kuru* (to come)
ima/genzai | *atari/hen ni*	*iru* (to be there)

5 Evaluation

To evaluate our methods, we employ ATR SLDB, which is our observation target in Section 3, as the trained text set and ATR LDB (Language DataBase) (Furuse et al., 1994) as the unseen text set. The two corpora were independently collected, but both have the same domain and task, i.e., travel type conversations between two persons. Table 5 shows the number of dialogs and utterances of both corpora.

Table 5: Two corpora: SLDB and LDB

corpus	SLDB	LDB
dialogs	618	1629
utterances	15425	35937

5.1 Evaluation Measure for Paraphrasing

Before reporting our evaluation results, we first need to discuss our evaluation measure for paraphrasing. To date, discussions on paraphrasing evaluations have been insufficient, since

Table 3: Examples of paraphrasing

	labeled utterance (input format)	applied pattern
0	おh支払いnのq␣方nはp␣**いかがd**␣なさいvますtかe	`s/いかがd/どうk/`
1	**おh支払いnのq**␣方nはp␣どうk␣なさいvますtかe	`s/おh支払n/支払n/`
2	支払いn**のq**␣**方n**はp␣どうk␣なさいvますtかe	`s/のq（方n|ほうn)//`
3	支払いnはp␣どうk␣**なさいv**ますtかe	`s/なさいv/しv/`
4	支払いnはp␣どうk␣しvますtかe	

paraphrasing research itself has practically just started.

We understand that we need to evaluate at least two measures for paraphrasing in general:

- accuracy: how good (correct/natural/...) paraphrases are produced.
- coverage: how many utterances are paraphrased correctly.

We should also consider a measure, unique for paraphrasing, that has the ability to determine the following: how many paraphrases are able to be produced. In this paper, however, we do not need to employ this measure, since we discuss a simplification by paraphrasing honorifics. This measure would be utilized if we were to discuss the generation of honorifics.

The two measures, the accuracy and the coverage, can be applied in evaluating the results of the disambiguation. This is because each disambiguation result is very clear regardless of whether it is correct or not, since the disambiguation method is based on deterministic rules and employs no learning method. Therefore, we evaluate our disambiguation rules by these two measures.

5.2 Evaluation of paraphrasing

We randomly pick out 1000 utterances from each corpus, and judge their accuracies. Table 6 shows results of the evaluation, and also shows the number of utterances (shown as ‘paraphrasable' in Table 6) that were not paraphrased but should have been paraphrased by the proposed method. The number of utterances that should have been paraphrased enables us to calculate the recall shown in Table 6. The coverage in the table shows the ratio of utterances that were paraphrased correctly to all utterances, while, the recall is the ratio to all utterances that should have been paraphrased.

The number of paraphrased utterances shown in Table 6 does not necessarily express how many honorifics were paraphrased. In addition, they include other non-honorifics that were paraphrased, such as phonemic changes.

Table 6: Evaluation results of paraphrasing

	observed	unseen
corpus	SLDB	LDB
utterances	1000	1000
paraphrased	745	770
unacceptable	1	8
paraphrasable	12	23
accuracy	99.9%	99.0%
coverage	74.4%	76.2%
recall	98.4%	97.1%

5.3 Evaluation of disambiguation

We randomly pick out 100 utterances, which have pre-determined ambiguous honorific verbs, from each corpus. In addition, each utterance has just one ambiguous honorific verb. We apply our disambiguation method, and evaluate the disambiguation results.

Table 7 shows results of the disambiguation, where R_{human} and R_{clue} are defined as follows:

$$R_{human} = \frac{f_{human}}{100 - f_{comp}} \quad (1)$$

$$R_{clue} = \frac{f_{clue}}{f_{human}} \quad (2)$$

where f_{human} represents the number of verbs that could be disambiguated by humans, f_{comp} represents the number of verbs that could be disambiguated automatically, and f_{clue} represents the number of verbs with obvious clues useful for automatic disambiguation in verbs that could be disambiguated by humans.

In other words, R_{human} represents how many verbs were disambiguated by humans in verbs that could not be disambiguated automatically, and R_{clue} represents how many verbs had obvious clues (which were able to be mounted on the disambiguation method) in verbs that could be disambiguated by humans.

6 Discussions

The evaluation results shown in Table 6 prove that the method could paraphrase almost 75% of the

Table 7: Disambiguation results

	seen	unseen
corpus	SLDB	LDB
utterances	100	100
disambiguated	66	70
coverage	66.0%	70.0%
accuracy	100%(66/66)	100%(70/70)
R_{human}	67.6%(23/34)	73.3%(22/30)
R_{clue}	39.1%(9/23)	63.6%(14/22)

utterances with a very high accuracy. We can therefore conclude that the paraphrasing method is very safe, namely, it has vanishingly scarce side effects.

From Table 7, we can state that the disambiguation method is also very safe. However, humans could not disambiguate almost 10% (11/100 in SLDB, 8/100 in LDB) of the utterances, and even if they could disambiguate them in some way, almost 10% (14/100 in SLDB, 8/100 in LDB) of the utterances had no clues for automatic disambiguation. From these observations, the possible coverage of the proposed disambiguation method is limited to almost 80%. The reason why humans could not disambiguate about 10% was the lack of information. In particular, the disambiguation of verbs, such as *irassyaru* (either to come, go, or be there) required information on the position between the speaker and his or her audience or a third person that was mentioned.

What is simplification by paraphrasing? Although we evaluated the paraphrased results on whether they were acceptable or not, unfortunately we did not take any objective measure to evaluate the simplification. However, if we were to introduce the number of chunks as a measure for such simplification, the results might be promising as shown in Table 8.

Table 8: Paraphrasing results of whole corpora

name	SLDB	LDB
utterances	15425	35937
utterances paraphrased	11643	27148
chunks	77035	176030
chunks paraphrased	71462	161797

In this work, we do not cover passive-formed honorifics. To cover them, we need to examine how to process them correctly and paraphrase them adequately in future work. As a stepping stone to paraphrasing passive-formed honorifics, we investigated how many passive-formed honorifics were included in SLDB. We counted the number of passive forms of "V + basic form of *(ra)rereu*" from SLDB parsed by JUMAN and KNP. In addition, we extracted sentences that included passive forms, and classified each passive form into three groups: used as an honorific, used as a passive form, or used as a potential verb form. Results of the investigation are shown in Table 9.

Table 9: Passive forms in SLDB

sentences in SLDB	15425
sentences including passive forms	407
passive forms	427
used as honorifics	241
used as passive forms	115
used as potential verb forms	49
could not classify	22

Table 9 shows that more than half of all passive forms are used as honorifics. A considerable part of the results we could not classify was caused by particular verbs: *omou* (to think), *kangaeru* (to consider), *kanjiru* (to feel), and so on. These verbs compose spontaneous passive sentences (Kaiser et al., 2001).

From a grammatical point of view, if a verb is a consonantal verb, then the passive form of the verb can not be a potential verb form but a really passive form or a subject-honorific form. This is because a consonantal verb has the original potential verb form "V + *eru*."

On the other hand, if a verb is a vocalic verb, the passive form of the verb can be one of the three types. Sometimes, however, the passive form of a vocalic verb omits '*ra*' if the form is a potential verb form. For example, *mirareru* (the passive form of *miru* (to see)) can be used for a really passive form or a subject-honorific form, and the form *mireru* dropping '*ra*' from *mirareru* can be used for a potential verb form.

From the discussion above, we may be able to narrow the candidates of a passive form down. However, to paraphrase passive forms correctly, there are some other obstacles that have to be solved:

- ellipsis resolution: In Japanese, ellipses are seen very often. As a consequence, if an utterance has a passive form, and the subject or object of the utterance is omitted, the passive form is hard to paraphrase.
- case frames: To determine whether a passive form is a really passive form or not, high quality case frames are strongly desired.

7 Related Works

To date, there had been no work directly related to the paraphrasing of honorifics. This is because the methods or techniques in SLT or other NLP applications had been premature. However, there are some slightly related works.

Maeda et al. (1988) proposed a unification-based approach to Japanese honorifics based on a version of HPSG (Head-driven Phrase Structure Grammar) for an experimental system that translates Japanese-English telephone and inter-keyboard dialogs. They also discussed that many human anaphoric references can be resolved by recourse to pragmatic constraints on the use of honorifics. Unfortunately, however, they did not discuss the domain or range of their approach, and they also did not evaluate it.

Siegel (2000) presented a solution for the representation of Japanese honorificational information in the HPSG framework for a machine translation system. The difference between Siegel's solution and ours is whether the method assumes a framework, such as HPSG, or not. From a practical viewpoint, our proposed method does not assume such a theoretical framework but only shallow methods such as POS tagging, dependency analysis, and pattern matching in *Perl*. That is to say, a considerable part in the paraphrasing of honorifics can be done by a string level pattern matching technique, where the method assumes only shallow level techniques like those mentioned above.

In the field of paraphrasing for simplification, Chandrasekar et al. (1996) discussed an approach of paraphrasing text by syntactical simplification, based on the Finite State Grammar and a supertagging model. Actually, they have the same motivation as us in believing that simplification is of great use for both humans and machines. However, their direction to simplification seems different from ours: they attempt to separate long and complicated sentences, whereas our target is to reduce variations spoken in the real world.

8 Conclusion

This paper discussed the paraphrasing of Japanese honorifics by manually constructed rules. Through an evaluation, it was unveiled that our simplifying method can paraphrase over 70% of all utterances with a very high accuracy (99%). It was also proved that the proposed disambiguation method is very safe. However, the possible coverage of this method is limited to almost 80%. To disambiguate beyond this limit of coverage, we need more information, such as the positions of the participants.

References

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and method for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 1041–1044.

Osamu Furuse, Yasuhiro Sobashima, Toshiyuki Takezawa, and Noriyoshi Uratani. 1994. Bilingual corpus for speech translation. In *Proceedings of AAAI '94 Workshop on the Integration of Natural Language and Speech Processing*, pages 84–91.

Stefan Kaiser, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto. 2001. *Japanese: A Comprehensive Grammar*. Routledge.

Hiroyuki Maeda, Susumu Kato, Kiyoshi Kogure, and Hitoshi Iida. 1988. Parsing Japanese Honorifics in Unification-based Grammar. In *Proceedings of the 26th Annual Meeting of the ACL*, pages 139–146.

Tsuyoshi Morimoto, Noriyoshi Uratani, Toshiyuki Takezawa, Osamu Furuse, Yasuhiro Sobashima, Hitoshi Iida, Atsushi Nakamura, Yoshinori Sagisaka, Norio Higuchi, and Yasuhiro Yamazaki. 1994. A speech and language database for speech translation research. In *Proceedings of ICSLP '94*, pages 1791–1794.

Melanie Siegel. 2000. Japanese Honorification in an HPSG Framework. In *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, pages 289–300.

Kazuhide Yamamoto, Satoshi Shirai, Masashi Sakamoto, and Yujie Zhang. 2001. SANDGLASS: Twin paraphrasing spoken language translation. In *Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL2001)*, pages 154–159.

David Yarowsky. 2000. Word-sense disambiguation. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 629–654. Marcel Dekker, Inc.