Anaphora Resolution for Transforming Regular Expressions into Honorifics in Japanese

Miao Yu, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, Japan
{miao-y,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

1. Introduction

The Japanese language places a heavy emphasis in establishing hierarchical relations among people, or paying respects to elders or those of higher social rank. The Japanese language includes the normal speech (常語) and honorific speech (敬語), with honorific speech playing an important role in establishing these relationships. The honorifics in Japanese include different levels of respectful (尊敬語), humble (謙譲語), and polite (丁寧語) speech, which are frequently used in various social or business situations. Because of this, one verb in Japanese will have at least four ways to express the same meaning in different situations. For example, word 行く (go), it will have the four expressions as in Table 1.

Table 1: Honorific Expressions

Honorific type	Word
Normal (基本形)	行く
Respectful (尊敬語)	いらっしゃいます
Humble (謙譲語)	参ります
Polite (丁寧語)	行きます

The mechanism of honorifics in Japanese is complicated, and many non-native Japanese speakers, as well as members of the young generations in Japan, have trouble mastering it. This situation has encouraged the study of automatic systems that identify the proper form of honorifics in Japanese including automatic translating of learners text into correct honorifics [1, 2]. However these previous works do not have a mechanism to determine which kind of honorific form should be applied in sentences without subjects. As subject omission is a common phenomenon in Japanese, this limits the practical utility of these methods.

In this paper, we propose the use of anaphora resolution for this task. We incorporate anaphora resolution into a rule-based machine translation system to translate regular expressions into appropriate honorific of Japanese, and examine the effectiveness of using correct subjects annotated by a human, and those automatically predicted by anaphora resolution.

2. Automatic Correction of Japanese Honorifics

2.1. Japanese Honorifics

The Japanese language has many honorifics, expressions which show respect, which are used in many social situations. The system of honorifics in Japan is very extensive, including respectful, humble, polite, and word beautification expressions. We show an overview in Table 2. Based on the relationship between the speaker and the subject of the sentence, there will be different honorific forms to be expressed. So, the subject of the sentence plays an important role in choosing an appropriate honorific in Japanese.

2.2. Related Work

There has been some related work on automatic processing of honorifics in Japanese. For instance, Noguchi et al. proposes a system to generate all varieties of honorific forms for single verbs automatically [1]. In their study, verbs are considered exclusively, and no contextual information is employed. In another study, Shirado et al. proposed a system for flexibly judging the misusage of honorifics in a Japanese sentence [2]. The system can point out misused words and phrases, and can also indicate how they are misused. The advantage of their research is that it can be applied in practical computer aided education. However, there is one major shortcoming of both works in that both of these works cannot deal with sentences for which the subject has been omitted. We investigate this issue further in Section 3.

3. Anaphora Resolution for Honorific Translation

3.1. Honorific Translation Task

Japanese honorifics are used in a wide variety of communication such as dialogue, interviews, meetings and emails. In this paper, we focus mainly on emails, as it is relative easy to collect sample data, and a system to correct polite expressions in email would be of great practical use to both foreign-

Table 2: Types of honorifics

Honorific type	Definition	Morphemes	Example
Respectful (尊敬語)	A special form or alternate word used when talking about superiors and customers.	Verbs with a higher-up as the subject; Nouns	いらっしゃいます、お返事
Humble (謙譲語)	Similar to respectful language, but used when talking to superiors about oneself.	Verbs with a lower rank person as the subject, usually the speaker; Nouns	お書きします、弊社
Polite (丁寧語)	Polite speech that doesn't fall into the above categories. Use of the copula "desu" and the verb ending "masu".	Verbs with the subject as friends or other people who's relationship is not clear; Nouns; Adjectives	帰ります、 美味しいです、本です
Word beautification (美化語)	Making words more polite or "beautiful". It is commonly achieved by adding the prefix o- or go- to a word.	Nouns	お茶

ers and native speakers.

3.2. Honorific Translation System

In our system, we focus on the correction of verbs, and divide the processing into four steps:

Identification of the Reader: As we are considering emails, in most cases the first line will consist of the addressee. We use this information to identify the reader of the mail. Next, we check the relative rank of the reader compared to the writer. We do this by checking if the writer uses titles that indicate higher rank including "先生、社長、部長、様". A total of 20 kinds of titles are as chosen higher-ranking titles. Titles not included in the higher lists, such as "さん" are designated as the same or lower rank than the writer.

Identification of Verbs: In this step, we used Mecab [3], to analyze every sentence in the document. Based on this result, we choose all verbs as potential candidates for correction.

Identification of Verb Subjects: In this step, we used the result of Mecab, to identify nouns in every sentence in the document. Then, we either recognize the first noun as the subject of each sentence in the baseline system, or use anaphora resolution as described in the next section.

Generation of the Correct Verb Form: In the final step, we divided the subjects into three categories. When we recognize the subject of sentence is "私、私たち、うち" and the reader has been identified as higher ranker we translate into the Humble form, in which the verb will be written as "お + conjunctive form + いたします". When we recognize the subject of the sentence is higher-ranking than the writer, we translate into Respectful forms, in which the verb will be written as "お + conjunctive form + になります". All other subjects are recognized to be Polite forms, in which the verb will be written as "conjunctive form + ます". Then, we choose the correct verb form based on the subject and context of each verb. Based on our analysis of the data, we also identified a number of instances that do not fit in the above framework. Those are added as rules as shown in Table 3.

Table 4: Different honorific expressions with different subjects

Subject	Listener	Sentence
higher	—	今日、大阪にいらっしゃいます。
self	higher	今日、大阪に参ります。
self	same / lower	今日、大阪に行きます。

3.3. Anaphora Resolution

As we explained in Section 3.1, based on different subjects, we will determine different honorific forms. However, Japanese is a pro-drop language, which means that the subject of the sentence can often be omitted. But for different subjects in the same sentence, there will be different expressions for the honorifics. For example, in the sentence "今日、大阪に行く。", based on the latent subject, we will use different expressions, as shown in Table 4. This shows that it is important to identify a subject in sentence before we translate it into appropriate honorific forms in Japanese.

Zero-anaphora resolution is a technology that has the ability to identify dropped subjects in these situations. For

Table 3: Additional rules to handle exceptions in honorific translation

Rules	
1	When the subject does not include the speaker and if the input has “しましょう”, the output will be changed into “します”.
2	When the subject is the speaker and if the input has “おる”, the output will be changed into “いる”.
3	If the sentence has “ますです” and the output will be deleted “です”.
4	If the sentence has “しよ、しましょ” and the output will be changed into “しよう” or “しましょう”
5	When the sentence has “ます、です、する” before “、” add “が” before “、”.
6	If the sentence has “よる、から、ので、” and the verb before will not be translated.
7	If the subject is of higher rank “よろしいでしょうか” will be translated into “いかがでしょうか”.

zero-anaphora resolution, Iida et al. approached this problem by decomposing it into intra-sentential and inter-sentential zero anaphora [4]. They improved the accuracy of intasentential zero-anaphora, which consequently improved the overall performance of zero-anaphora resolution. In another study, Yeh et al. used shallow parsing instead of complex parsing to resolve zero anaphora. In their study, the precision rates of zero anaphora detection and the recall rate of zero anaphora resolution with the method are 81% and 70% respectively [5]. In our system, we choose to adapt the method of Iida et al. as it provides high performance and is available open source in the SynCha package [6].

4. Experiment

In the experiment, we evaluated the performance of the method proposed in Section 3.

4.1. Experiment Setting

As data for our experiment, we collect 6 emails written by non-native speakers as the source data for the proposed method. For the reference data, we asked a native speaker of Japanese well-versed in business to correct the source data into proper and polite Japanese.

We divided the source data into 2 types. The first type (Source) is regular source data which has mistakes of all varieties including verbs, nouns, and others. The other type (KeigoOnly) is source data that is modified manually so that the only mistakes are misuse of polite expressions. We do this because our system focuses on translating into appropriate honorifics of Japanese, so the second setting removes the effect of mistakes that are not the target of our system. We prepare three types of systems to compare:

No Resolution: translate the source data without resolving dropped subjects

SynCha: automatically resolve omitted subjects using SynCha and consider them when translating into honorifics

Human: annotate omitted subjects by hand and then consider them when translating into honorifics

Table 5: WER results on Source (WER is 36.05%)

	No Resolution	SynCha	Human
+Res+Hum	35.88%	35.53%	35.70%
+Res-Hum	35.53%	35.18%	**35.01%**
-Res+Hum	36.05%	36.22%	36.40%
-Res-Hum	35.53%	35.53%	35.53%

We also evaluated four different options regarding which forms to translate.

+Res+Hum: translate into appropriate respectful, humble, and polite expressions

+Res-Hum: only translate into respectful and polite expressions

-Res+Hum: only translate into humble and polite expressions

-Res-Hum: only translate into polite expressions

We measure the accuracy of our method using word error rate (WER) using the corrected text as a reference.

4.2. Evaluation and Results

First, we show the results of the experiment using the Source data as input in Table 5. The WER of Source is 36.05%. Table 5 shows that almost all forms of the system achieve better performance than the original source. Comparing accuracy of No Resolution, SynCha, and Human, we confirm that first resolving the omission of subjects, and then translating into appropriate honorifics of Japanese is useful and necessary. In addition the WER of the systems that generate humble expressions is higher than those that do not. This is because it is common to use humble forms only in some occasions such as when performing an action for a superior.

We also show the results of the experiment using the KeigoOnly data as input in Table 6. The WER of KeigoOnly is 15.25%. This Table 6 shows that the tendency is similar to when considering all errors.

Table 7: Examples of each system's output, with the previous sentence of context

Context	社長がお参加になりますか？
No Resolution	日本料理店でよろしいでしょうか？
SynCha / Human	日本料理店でいかがでしょうか？
Context	今年４月に大連理工大学を卒業した、素晴らしい技術を勉強するために、日本に留学来ました。
No Resolution / SynCha	先生のウェーブページを見て、ご研究の音声翻訳が気になります。
Human	先生のウェーブページを拝見して、ご研究の音声翻訳が気になります。

Table 6: WER results on KeigoOnly (WER is 15.25%)

	No Resolution	SynCha	Human
+Res+Hum	14.73%	14.38%	14.38%
+Res-Hum	14.38%	14.04%	**13.86%**
-Res+Hum	14.90%	14.90%	15.08%
-Res-Hum	14.38%	14.38%	14.38%

4.3. Discussion

We show some examples of the system output in Table 7. In the first, because the preceding sentence has the subject "社長", SynCha and the Human annotator recognize the subject of "よろしい" to be "社長". Thus, both are able to translate the "よろしいでしょうか" into "いかがでしょうか". However, in the second example, because the context does not directly include the subject "私", SynCha was not able to identify the subject appropriately, resulting in failure to translate into the appropriate honorifics. Identifying anaphora that refer to the reader or writer of a passage is a difficult task, but Hangyo et al. have recently proposed a method that may solve this problem [7].

It should be noted, however, that the number of corrections made by our model are somewhat small compared to the total number of phenomena to be corrected in KeigoOnly. One reason for this is that there are often multiple ways to express verbs politely in Japanese. For example, in some places our annotator changed "なりたいです" into "なりたいと思います". There are many such examples that our system cannot handle at this time.

Finally, nouns, adjectives, and other expressions also have honorific forms. Because we are only translating verbs, it is not sufficient to translate all of the honorific expressions in our corpus.

5. Conclusion

We proposed a method to incorporate anaphora resolution with a rule-based system to translate the regular expressions into appropriate honorifics of Japanese. Experimental results showed that this improved the performance of the honorific translation system.

In future research, we plan to expand to varieties of morphemes other than only verbs. We will also improve the translating performance of humble expressions.

6. References

[1] 野口聡, 南條浩輝, and 吉見毅彦. 動詞の通常表現から敬語表現への換言. In 言語処理学会第 *13* 回年次大会 *(NLP2007)*, 2007.

[2] Tamotsu Shirado, Satoko Marumoto, Masaki Murata, and Hitoshi Isahara. System for flexibly judging the misuse of honorifics in Japanese. In *Proceedings of Pacific Asia Conference on Language, Information, and Computation*, pages 503–510, 2011.

[3] Taku Kudo. MeCab: yet another part-of-speech and morphological analyzer. 2006.

[4] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 625–632, 2006.

[5] Ching-Long Yeh and Yi-Chun Chen. Zero anaphora resolution in Chinese with shallow parsing. In *Journal of Chinese Language and Computing*, pages 41–56, 2007.

[6] Ryu Iida and Massimo Poesio. A cross-lingual ILP solution to zero anaphora resolution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

[7] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Japanese zero reference resolution considering exophora and author/reader mentions. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 924–934, 2013.