Source Language Effect on Translating Korean Honorifics

Kyonghee Paik[1], Kiyonori Ohtake[1], Francis Bond[2],
Kazuhide Yamamoto[1,3]

[1] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City", Kyoto 619-0288 JAPAN
{kyonghee.paik, kiyonori.ohtake}@atr.co.jp

[2] Nippon Telegraph and Telephone Corporation
2-4 Hikaridai, "Keihanna Science City", Kyoto 619-0237 JAPAN
bond@cslab.kecl.ntt.co.jp

[3] Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka City, Niigata 940-2188 JAPAN
yamamoto@fw.ipsj.or.jp

Abstract. This paper investigates honorific phenomena on two variants of Korean translation corpus, based on translations from Japanese and English. One surprising result is how different the corpora were, even after normalizing orthographic differences. Translations are dependent not just meaning, but also on the structure of the source text.

1 Introduction

This paper investigates the effect of source language on honorifics in translations using two variants of a Korean translation corpus. The original corpus was compiled from Japanese-English parallel sentences collected from phrase books for Japanese traveling abroad. The corpora used for this research consist of 324,616 Korean sentences. Half of the Korean sentences (162,308 sentences: henceforth, K_J) were translated from Japanese and the other half (henceforth, K_E) have been translated from English sentences which match the Japanese.

Although the two Korean corpora are generally equivalent in meaning, they have different characteristics, since they were translated from such different languages as English and Japanese. We show that the differences between English and Japanese lead to different results in the Korean translation, even though the original source texts were matching Japanese-English translation pairs.

Our ultimate goal is to find the differences in translations by comparing two variants with different source texts in order to improve the quality of machine translation. So far such comparison has not been extensively studied.

2 Honorifics in Japanese and Korean

Korean is an honorific language with an extremely systematic grammaticalization. A sentence cannot be uttered without the speaker's knowledge of their social relationship to the addressee and referent considering social status, age, kinship, familiarity and so on. Otherwise, the utterance may sound rude, inappropriate, or even awkward (Sohn 1999, p.16).

Japanese honorifics are also expressed by various forms according to the degree of honor or respect, addressee, or situation. Like Korean, Japanese can make nouns and verbs honorific using different lexical items or adding prefixes and suffixes. For verbal constructions, Japanese uses *o/go- V ni naru* "do" (subject honorific) and *o/go -V suru* "do" (addressee honorific (humble)) as well as two different speech levels of verb endings: *da* (plain) *desu* (polite) (Kaiser et al 2001). Korean uses verbal suffixes *-si* to make subject honorifics and *-sup* to make addressee honorifics, along with six different speech levels.

There are many mismatches between the two systems from Japanese to Korean. For example, Japanese beautifies nouns *o-* or *go-* before nouns like *o-saifu* "wallet", *o-sio* "salt", *go-tsugo* "convenience". However, Korean does not have a corresponding system to this. As for verb honorifics at speech levels, Korean has six different levels whereas Japanese has only two levels. This will cause mistakes in machine translation between the two languages.

3 Analysis of our corpora

We compared sentences using the perl module `String::Similarity` (Lehman 2000). It returns a similarity score based on the edit distance (the number of characters that need to be deleted, added or substituted to change one string into another), normalized to give a score between 0 and 1 (see Meyers (1986) for a fuller description). Two completely different strings have a score of zero, while two identical strings have a score of one. There were 136,529 sentences. Their distribution is given in Table 1. As there were many identical sentences in each corpus, we give the distribution over both all sentences and unique sentences. Less than 2% of the sentences were the same in both corpora. Most sentences were reasonably similar (0.4–0.8). Very few sentences were identical (less than 4%) and even fewer were totally dissimilar (less than 0.1%). There were some minor orthographic differences. When they were resolved, only 8.3% of the sentences were identical. To put this in perspective, Culy and Riehemann (2003) discuss the fact that there is considerable variation in translations from exactly the same source text. It is impossible to explicitly compare their results for two reasons: (1) they are looking at more literary texts, where the translators certainly were aware of the previous translations and trying to differentiate their own; (2) they do not quantify the differences using the same measure as us.

We took one percent of the unique sentences for each of 9 similarity bands from 0.1–0.2 to 0.9–1.0 for more detailed evaluation. The sample size was proportional to the number of sentences in each band — so the samples are small at each extreme and large in the center. We evaluated 1,360 randomly chosen sample sentences and all 58 sentences of the least similar band (0.0–0.1).

Since our corpora deals with the topics related to various travel situations, there are numerous sentences inquiring and trying to get information which are

Table 1. Distribution of the sentences by similarity score

	Similarity Score										
	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	Total
S_{all}	100	1,910	11,006	23,126	33,755	34,888	28,083	17,400	7,693	4,347	162,308
S_{uniq}	58	1,243	7,876	19,351	29,053	30,149	24,382	14,946	6,434	3,037	136,529

uttered using polite or deferential forms of interrogatives in most cases. The different use of adjectives, adverbs, or their phrases are also found all through different similarity scores. These phenomena form a useful source for learning lexical paraphrasing rules for example-based paraphrasing.

We found strikingly different distributions in K_J and K_E with regard to honorifics, which are shown in Table 2 and discussed below.

Table 2. Honorific differences between K_J and K_E

	Similarity Score									
	0.0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
Samples	58	12	78	192	290	300	243	149	64	30
Honorific	(K_J) 4	1	10	16	82	101	76	32	11	2
Honorific	(K_E) 2	0	8	15	20	32	31	9	2	0
Loan word	(K_J) 0	0	5	16	14	15	7	6	0	0
Loan word	(K_E) 0	0	0	7	11	9	11	3	0	0

The differences shown in Table 2 mainly arise from the fact that K_J tends to use more deferential honorifics whereas the K_E tends to use more polite honorifics, although this distinction is not made in either English or Japanese. Even so, the source language affects the degree or level of honorifics. For example, K_E tend to use polite speech levels like -요 *-yo* where as K_J tend to use deferential forms like -습니까 *-(su)pnika* for asking questions. This systematic difference makes these two corpora a suitable source for the automatic acquisition of paraphrasing rules. The difference in honorific use is strongly shown in the bands from 0.4 to 0.9. In general, K_J used more honorifics than K_E overall, and the effect was to make it a more natural collection of travel phrases.

As we have seen, the source language has a large effect on the translation, although we have focused on honorifics in this paper. This suggests that the optimal translation strategy may be different between language pairs. In particular, a system translating between Japanese and Korean needs to put less effort into lexical and syntactic choice, and more into the use of honorifics. A system going between English and Korean has a much harder task, and must consider lexical and syntactic choice, zero pronoun resolution in addition to the use of honorifics.

Differences in honorific usage appeared to be fairly predictable between the two corpora, and lead to the hope that a paraphrasing module can be used to fix the honorific levels after translation itself occurs, along the lines suggested by Ohtake and Yamamoto (2001).

Another interesting difference was in the distribution of borrowed words or foreign words from languages other than Chinese. We expected that more loan words would be used in K_E because most loan words come from English. However, according to Table 2 the result is the opposite. Further, the loan words translated from Japanese, were in general more natural (73%) than those translated from English (50%). After examining the source corpora of K_J, we found that many "Katakana" loan words are used in the Japanese corpus. Almost all of the Katakana words are translated into loan words in Korean. In contrast, all the English words are equally foreign, so for any word, there is little pressure for it to be translated into a loan word, rather than into native Korean. When English words were translated as loan words, they tended to be poor literal translations.

4 Conclusion and Further Work

We investigated two variants of a Korean translation corpus, based on translations from Japanese and English. We have shown that the source language text has a large influence on the target text. One surprising result is how different the corpora were, even after normalizing orthographic differences. In practice, translations are dependent not just on meaning, but also on the structure of the source text. We find that the expressions on honorifics which were examined using similarity scores are reliable resources for paraphrasing.

Based on these findings, we intend to examine whether we can automatically extract grammatical, semantic and lexical paraphrases by comparing corpora using similarity scores.

Acknowledgement

This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

Culy, C. & S. Z. Riehemann: 2003, 'The limits of N-gram translation evaluation metrics', in *Proceedings of MT Summit IX*, New Orleans.

Kaiser, Stefan, Y. Ichikawa, N. Kobayashi & H. Yamamoto: 2001, *Japanese: A Comprehensive Grammar*, Routledge.

Lehmann, Marc: 2000, 'String::Similarity', Perl Module (cpan.org), (v0.02).

Myers, Eugene: 1986, 'An O(ND) difference algorithm and its variations', *Algorithmica*, **1**(2): 251–266.

Ohtake, Kiyonori & Kazuhide Yamamoto: 2001, 'Paraphrasing honorifics', in *NLPRS-2001*, Tokyo, pp. 13–20.

Sohn, Ho-Min: 1999, *The Korean Language*, Cambridge Language Surveys, Cambridge University Press.