Introducing Translation Dictionary Into Phrase-based SMT

Hideo OKUMA, Hirofumi YAMAMOTO and **Eiichiro SUMITA**
NICT Spoken Language Group
ATR Spoken Language Communication Laboratories
Hikaridai 2-2-2, Seika-cho, Soraku-gun, Kyoto, Japan
{hideo.okuma, hirofumi.yamamoto, eiichiro.sumita}@nict.go.jp

Abstract

This paper presents a method to effectively introduce translation dictionaries into phrase-based SMT. Though SMT systems can be built with only a parallel corpus, translation dictionaries are more widely available and have many more entries than parallel corpora. A simple and low-cost method to introduce a translation dictionary is to attach a dictionary entry into a phrase table. This, however, does not work well. Target word order and even whole target sentences are often incorrect. To solve this problem, the proposed method uses high-frequency word in the training corpus. The high-frequency words may already be trained well, in other words, may appear in the phrase table and therefore be translated with correct word order. Experimental results show the proposed method as far superior to simply attaching dictionary entries into phrase tables.

Introduction

Statistical Machine Translation (SMT) systems are built solely based on a large parallel corpus. The performance of SMT has improved by using "phrase" units for translation, rather that "word" units. With phrase-based SMT (Koehn et al., 2003), the term "phrase" is used to mean a sequence of words, as opposed to a linguistical phrase. Within the phrase, the selection of target words and the order of target words are learned in advance and left unchanged during the translation process.

SMT systems are trained using only a parallel corpus. SMT cannot translate unknown words, words that do not appear in the parallel corpus. Furthermore, unknown words often destroy the entire translation, resulting in a sequence of scattered words that is beyond comprehension. There are many translation dictionaries available and they are a more common resource than parallel corpora. There are general dictionaries, such as EIJIRO[1] and EDICT[2]. Additionally, there are dictionaries with more technical/specialized terminology - dictionaries for patent, engineering, medical, legal, sport, entertainment, et cetera, such as LSD[3].

These are considered dependable because they have evolved not by using a fully automated process (which is exploited in the SMT paradigm), but by long-term human effort. It is reasonable to incorporate these valuable resources into SMT. We cannot expect a parallel corpus to include all necessary words. Even the publicly available and largest parallel corpus for the NIST MT competition (consisting of 8 million Chinese and English sentences) does not include many names of places and people. A mechanism is needed for handling words unseen in the parallel corpus. This paper puts focus on proper nouns because they are typically not found in the training corpus.

[1]http://www.alc.co.jp/
[2]http://csse.monash.edu.au/ jwb/j_edict.html
[3]http://lsd.pharm.kyoto-u.ac.jp/en/

Phrase-based SMT

Here, we explain phrase-based SMT, which is now regarded as the de facto standard. Before we explain phrase-based SMT, however, we shall briefly describe word-based SMT (Brown et al., 1993). This system is based on the noisy channel model. According to Bayes' law, the translation probability for translating source sentence f into target sentence e is represented as

$$argmax_e p(e|f) = argmax_e p(f|e) \times p(e) \quad (1)$$

$p(f|e)$ represents the translation model and $p(e)$ represents the language model. Whereas word-based SMT bases translations word-by-word, phrase-based SMT is based on a phrase-by-phrase translation model. In phrase-based SMT as discussed in this paper, the right-hand side of equation 1 is as follows:

$$argmax_e p_\phi(f|e) \times p_{LM}(e) \times p_D(e,f) \times \omega^{length(e)} \quad (2)$$

where $p_\phi(f|e)$ is a phrase translation model, $p_{LM}(e)$ is a language model, $p_D(e,f)$ is a distortion model and $\omega^{length(e)}$ is a word penalty. These are weighted. The translation process of phrase-based SMT is as follows:

1. Segment the source sentence into phrases

2. Translate the source phrases in any order stochastically

3. Adjust the position of the target phrases stochastically

Figure 1 shows phrase-based SMT with source phrases translated into target phrases. In each phrase, the word order is correctly maintained. All of the phrases in the translation process appear in a phrase table. The phrase table is a translation model for phrase-based SMT and consists of source language phrases and corresponding target language phrases and these probabilities. Figure 2 shows an example of a phrase table.

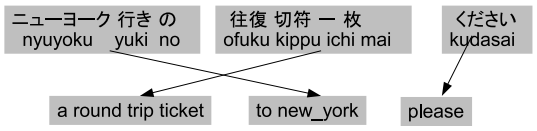

Figure 1: Phrase-based SMT

Baseline Method

A simple and low-cost method to introduce a translation dictionary into phrase-based SMT is to add parallel word pairs in the translation dictionary to the phrase table with appropriate probabilities. This method, however, does not work well. Although words are translated correctly, positions of translated words are not always correct. This is very serious, especially for language pairs in which word order is very different, such as Japanese and English. Figure 3 shows how this method works in a case where the source sentence is the same as that used in figure 1, but the Japanese word "ニューヨーク/nyuyoku (new york)" is replaced by the untrained Japanese word "カーディフ/kadifu (cardiff)" (a local place name in Wales, UK).

In this example, the Japanese word "カーディフ/kadifu" is translated to the English word "cardiff" correctly, but its position and the entire sentence become incorrect. This is because the Japanese word "カーディフ/kadifu" is not included in any source

カーディフ 行き の 往復 切符 一 枚 ください
kadifu yuki no ofuku kippu ichi mai kudasai
↓
cardiff for a round trip ticket please

Figure 3: Translation example of the baseline method

phrases of the phrase table apart from the one we just added, even though constraint of word order in phrase-based SMT deeply depends on the phrase itself. The language model which controls word order also does not include the English word "cardiff" and so cannot decide the word position correctly.

Proposed method

Basic idea

Figure 4 shows the process of the proposed method by example.

Instead of using the baseline method which has the above problem, we propose a method which uses the high-frequency words in the training corpus. Prior to the translation, the untrained words in source sentences are replaced with high-frequency words in the training corpus. The target sentences are then acquired by translating the modified source sentences. Finally, the high-frequency words in the target sentences are replaced with target words for the untrained words. The reason why we use high-frequency words is that they may already be trained well, in other words, the high-frequency words may already appear frequently in phrase tables and therefore provide ample statistics. It is also important that the high-frequency word be of the same category as the untrained word. By using high-frequency words of the same category, the contexts of both the untrained words and the high-frequency words are usually the same. The modified source sentence with the high-frequency word is then translated as the original source sentence. Using figure 4 , we describe the process step by step. First, the untrained word "カーディフ/kadifu" is replaced in the

```
ニューヨーク 行き の ||| to new_york |||  0.00898204 0.00202775 0.75 0.128232 2.718
(nyuyoku yuki no)
ニューヨーク 行き の 列車 ||| train to new_york ||| 0.666667 6.27302e-05 0.5 0.0419405 2.718
(nyuyoku yuki no ressya)
ニューヨーク 行き の 切符 ||| ticket to new_york ||| 0.25 6.32731e-05 1 0.034494 2.718
(nyuyoku yuki no kippu)
.....
往復 切符 ||| a round trip ticket ||| 0.422764 0.105845 0.675325 0.0743079 2.718
(ofuku kippu)
往復 切符 を 一 枚 ||| a round trip ticket ||| 0.0493827 9.23938e-06 0.5 0.006118 2.718
(ofuku kippu wo ichi mai)
往復 切符 を 一 枚 ください ||| one return ticket please ||| 1 6.42418e-07 0.333333 0.0018579 2.718
(ofuku kippu wo ichi mai kudasai)
```

Figure 2: A example of the phrase table for Japanese to English translation

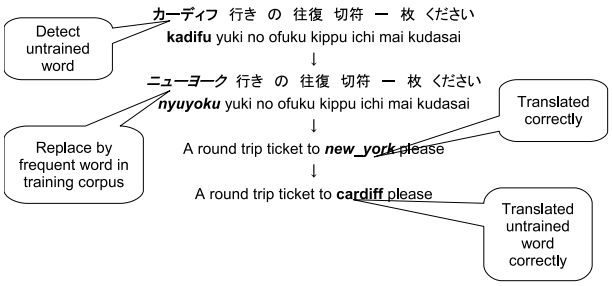

Figure 4: Translation process of the proposed method

source sentence with the high-frequency and well-trained word, "ニューヨーク/nyuyoku". Both "カーディフ/kadifu" and "ニューヨーク/nyuyoku" are of the same category "place-name". Next, the entire modified source sentence is translated and the target sentence is acquired. Next, the target sentence is searched for the translated word "new_york" . Finally, the high-frequency word "new_york" is replaced with "cardiff".

Formal explanation

Using high-frequency words assumes that the untrained word and the high-frequency word will have the same context once replaced, as both the previous word and the following word remain the same. So, in the case of a language model, this is a trigram:

$$p(w_{i-2}, w_{i-1}|w_{OOV}) = p(w_{i-2}, w_{i-1}|w_{freq}) \quad (3)$$

$$p(w_{i+1}, w_{i+2}|w_{OOV}) = p(w_{i+1}, w_{i+2}|w_{freq}) \quad (4)$$

where w_k are words in context, w_{OOV} is the untrained word and w_{freq} is the high-frequency word.

From equation 3,

$$\frac{p(w_{i-2}, w_{i-1}, w_{OOV})}{p(w_{OOV})} = \frac{p(w_{i-2}, w_{i-1}, w_{freq})}{p(w_{freq})}$$

$$\frac{1}{p(w_{OOV})} \times \frac{p(w_{i-2}, w_{i-1}, w_{OOV})}{p(w_{i-2}, w_{i-1})} = \frac{1}{p(w_{freq})} \times \frac{p(w_{i-2}, w_{i-1}, w_{freq})}{p(w_{i-2}, w_{i-1})}$$

$$\frac{1}{p(w_{OOV})} \times p(w_{OOV}|w_{i-2}, w_{i-1}) = \frac{1}{p(w_{freq})} \times p(w_{freq}|w_{i-2}, w_{i-1})$$

$$p(w_{OOV}|w_{i-2}, w_{i-1}) = \frac{p(w_{OOV})}{p(w_{freq})} \times p(w_{freq}|w_{i-2}, w_{i-1})$$

Because $p(w_{OOV})/p(w_{freq})$ is a constant, the language model scores of both $p(w_{OOV}|w_{i-2}, w_{i-1})$ and $p(w_{freq}|w_{i-2}, w_{i-1})$ are same up to constant factors.

Likewise, from equation 4,

$$\frac{w_{OOV}, p(w_{i+1}, w_{i+2})}{p(w_{OOV})} = \frac{w_{freq}, p(w_{i+1}, w_{i+2})}{p(w_{freq})}$$

$$\frac{p(w_{OOV}, w_{i+1})}{p(w_{OOV})} \times \frac{p(w_{i-2}, w_{i-1}, w_{OOV})}{p(w_{OOV}, w_{i+1})} = \frac{p(w_{freq}, w_{i+1})}{p(w_{freq})} \times \frac{p(w_{i-2}, w_{i-1}, w_{freq})}{p(w_{freq}, w_{i+1})}$$

$$p(w_{i+2}|w_{OOV}, w_{i+1}) \times p(w_{i+1}|w_{OOV}) = p(w_{i+2}|w_{freq}, w_{i+1}) \times p(w_{i+1}|w_{freq})$$

$$p(w_{i+2}|w_{OOV}, w_{i+1}) = \frac{p(w_{i+1}|w_{freq})}{p(w_{i+1}|w_{OOV})} \times p(w_{i+2}|w_{freq}, w_{i+1})$$

Because, by assumption, $p(w_{i+1}|w_{freq})/p(w_{i+1}|w_{OOV})$ is 1, the language model scores of both $p(w_{i+2}|w_{OOV}, w_{i+1})$ and $p(w_{i+2}|w_{freq}, w_{i+1})$ are equal.

Algorithm

Figure 5 shows the overview of the proposed method. As preparation, categorize the untrained translation dictionary into a number of categories such as place, person, organization and so on. Next, collect high-frequency words from each category from the training corpus and attach the translated word to each collected high-frequency word, and last, make a "surrogate list" which keeps all of the above in pairs. A surrogate list is prepared for each category. Here is the algorithm which replaces the untrained word in a source sentence.

1. For each word in a source sentence
2. For each category in all categories
3. If a word is not in a dictionary of a category, go to 6
4. Acquire the frequent word from a surrogate list of this category and replace the word in the source sentence with this high-frequency word.
5. Keep a tuple of the untrained word and its replaced word pair in the replaced table.
6. Repeat 2
7. Repeat 1

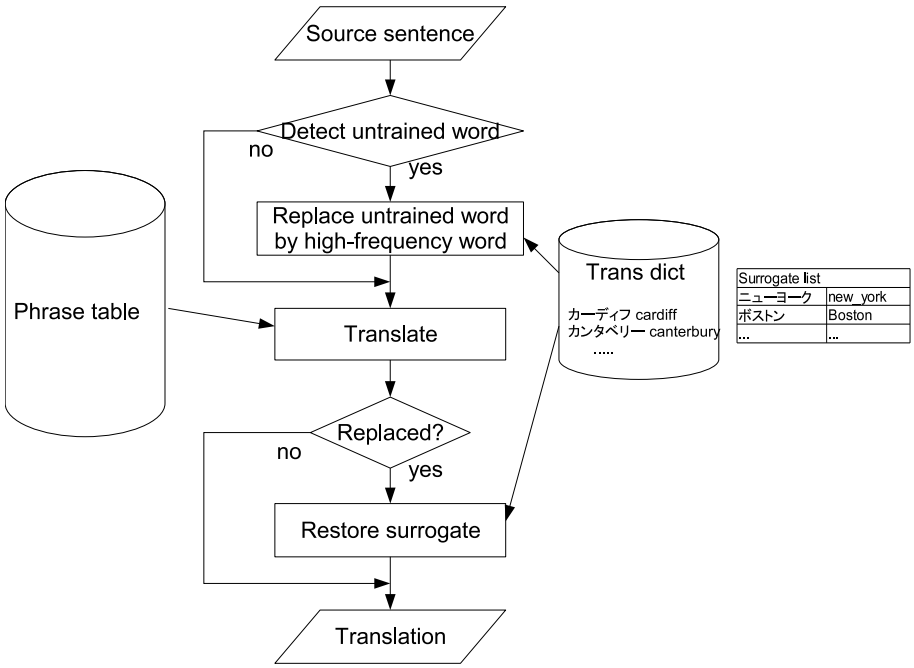

Figure 5: Overview of the proposed method

		Japanese	English
Training set	no. of sentences	681401	
	no. of tokens	5742797	4831057
	ave. of tokens	8.43	7.09
Test set	no. of sentences	10150	
	no. of tokens	74778	61224
	ave. of tokens	7.37	6.03

Table 1: Corpus size

In step 3, when acquiring a high-frequency word from a surrogate list, check whether the high-frequency word is already in the source sentence. If the high-frequency word is in source sentence, skip this high-frequency word and acquire the next one. In step 4, a result of all replacements is kept in the replaced table. This table is used to search which untrained word is replaced with which high-frequency word.

Use top N high-frequency words

Use of the top high-frequency word may suffice if the training corpus is sufficiently large. However, if the training corpus is not so large, use of the top N($N \geq 2$) high-frequency word may prove better than the top high-frequency word. This is because the top high-frequency word may not necessarily be trained well under such a training corpus. That is to say, the top high-frequency word may not appear in the phrase table with the same context of the untrained word, but the top N high-frequency word does. And it can also be said that regardless of training corpus size, it is better to try all of the candidates in a surrogate list to compare the results. We have also conducted an experiment to confirm this.

Experiment

Setting

For experiment, we use ATR's BTEC(Basic Travel Expression Corpus) (Kikui et al., 2003). This corpus contains basic expression for travel. Table 1 shows the size of this corpus.

We tagged the Japanese corpus using Chasen[4] and the English corpus using an in-house tagger at ATR and made both translation and language models for SMT system with these. These models were created using the training toolkit (Koehn and Monz, 2006) with GIZA++ (Och and Ney, 2003), mkcls (Och, 1999) and SRILM (Stolcke, 2002).

We made two special test sets from the original test set to clarify the differences between the proposed method and the baseline method. First, sentences were collected from the Japanese test set which included words tagged as "place-general" by Chasen. Those words were then replaced with the

[4]http://chasen-legacy.sourceforge.jp/

```
カーディフ ||| cardiff ||| 1.e-25 1.e-25 1.e-25 1.e-25 1.e-25
(kadifu)
ポーツマス ||| portsmouth ||| 1.e-25 1.e-25 1.e-25 1.e-25 1.e-25
(potsumasu)
ベルファスト ||| belfast ||| 1.e-25 1.e-25 1.e-25 1.e-25 1.e-25
(berufasuto)
カンタベリー ||| canterbury ||| 1.e-25 1.e-25 1.e-25 1.e-25 1.e-25
(kantaberi)
.....
```

Figure 6: A example of additional phrase table for the baseline method

untrained words “カーディフ/kadifu (cardiff)”, “ポーツマス/potsumasu (portsmouth)” and “ベルファスト/berufasuto (belfast)”. Second, sentences from the Japanese test set were collected, including words tagged as “person-name given” and “person-name family” and were then replaced with untrained words “ナオミチ (naomichi)”, “ミツアキ (mitsuaki)” and “ノブキチ (nobukichi)” for given and “コイズミ (koizumi)”, “オザワ (osaza)” and“ナカソネ (nakasone)” for family. We named these new test sets place name test set and person name test set. An English test set was prepared corresponding to these new Japanese test sets and replaced words were the same as the new Japanese test sets. These new English test sets were used as reference sets for the Japanese to English translation experiment. These test sets are also used as test sets for the English to Japanese translation experiment.

In the baseline method, we add these untrained replaced word pairs to a phrase table with the same format as described above. Figure 6 shows an example of additional word pairs. In the proposed method, we set the top five high-frequency words to the surrogate list for each of the categories.

Also, to ensure the advantage of the use of the top N high-frequency words, five modified source sentences are made by replacing the untrained words with the top five high-frequency words from the surrogate lists and the best translation result is chosen comparing scores as shown in equation 2. This is done if the untrained word is the only one in the source sentence to avoid any complexity of calculation.

We use Cleopatra made at ATR for the decoding, which is compatible with Pharaoh.

Result

Table shows results of an automatic evaluation. Values are scores of BLEU (Papineni et al., 2002). As the table indicates, the proposed method outperforms the baseline method with exception to the English to Japanese place name test set.

However, these BLEU scores are obtained by using only one reference set and have wide fluctuation, in contrast to using multiple references. To investigate further, differences between the translations produced by the two methods were checked using human evaluators.

Testset		place	person
	no. of sentences	106	60
JE	Baseline	39.87	36.87
	Proposed	**44.14**	**40.18**
EJ	Baseline	**42.09**	21.87
	Proposed	41.91	**25.53**

Table 2: Automatic evaluation(baseline vs. proposed)

Table shows the results of human evaluation. Human evaluation ranks are from A to D (A:perfect, B:fair, C:acceptable and D:nonsense) and we compare the percentage of A, A+B and A+B+C. The result of the human evaluation shows the superiority of the proposed method clearly. Table and shows both examples with ranks of the two methods. Table shows the result of human evaluation of those with only one difference: the top high-frequency word results versus the top 5 high-frequency word results. The table shows that the use of the top 5 method seems to be better than the use of the top 1 method, we count how many the use of top 5 translation results is better than the use of top 1 translation results and vice versa. Table shows this only if the rank is different. By sign test, the use of top 5 high-frequency words method is 5% significance than top 1’s.

Conclusion and future work

In this paper, we proposed a method to introduce a translation dictionary into a phrase-based SMT system and quantified the improvements.

In order to do this, we used a special test set to assess the differences between the proposed method and the baseline method. As a consequence, the test set became smaller than the original test set and in future research we would like to run experiements with a larger test set and/or multiple references.

In the case of inability to prepare a translation dictionary, we need to incorporate Named Entity recognition (Grishman and Sundheim, 1996) and translation/transliteration (Knight and Graehl, 1997) into the proposed method.

Also, this paper proposed our method only for certain categories of proper noun. For more general words, such as nouns or adjectives, we have a plan to apply our method. To this end it is important to categorize these general words automatically and to create a surrogate list for each category automatically.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The

Testset		place			person		
JE	no. of sentences	51			39		
	rank	A	A+B	A+B+C	A	A+B	A+B+C
	baseline	47.06%	72.55%	92.16%	38.46%	61.54%	82.05%
	proposed	**70.59%**	**84.31%**	**94.12%**	**64.10%**	**82.05%**	**89.74%**
EJ	no. of sentences	57			34		
	rank	A	A+B	A+B+C	A	A+B	A+B+C
	baseline	33.33%	75.44%	80.70%	50.00%	67.65%	91.18%
	proposed	**73.68%**	**80.70%**	**89.47%**	**58.82%**	**76.47%**	91.18%

Table 3: Human Evaluation(baseline vs. proposed)

Testset		place			person		
JE	no. of sentences	16			24		
	rank	A	A+B	A+B+C	A	A+B	A+B+C
	top 1	62.50%	**81.25%**	100.0%	50.00%	79.17%	87.50%
	top 5	**68.75%**	75.00%	100.0%	**70.83%**	**83.33%**	**91.67%**
EJ	no. of sentences	23			4		
	rank	A	A+B	A+B+C	A	A+B	A+B+C
	top 1	69.57%	82.61%	95.65%	50.00%	**100.0%**	**100.0%**
	top 5	**73.91%**	**91.30%**	95.65%	**75.00%**	75.00%	75.00%

Table 4: Human Evaluation(top 1 vs. top 5)

top 1	C	C	B	B	A	B	A	B	B	B	B	B	C
top 5	B	A	C	A	C	A	C	A	A	A	A	A	B
−/+	+	+	−	+	−	+	−	+	+	+	+	+	+
top 1	D	B	A	C	B	C	A	D	B	B			
top 5	C	A	D	A	A	B	B	B	D	A			
−/+	+	+	−	+	+	+	−	+	−	+			

Table 5: Differently ranked cases(top 1 vs. top 5 detail)

mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference- 6: A brief history. In *COLING*, pages 466–471.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proceedings of EuroSpeech 2003*, pages 381–384.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *ACL35TITLE*, pages 128–135, Madrid. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 127–133, Edmonton, Alberta, Canada, May 27 - June 1. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *EACL99*, pages 71–76, Bergen.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *ICSLP02*, pages 901–904.

Source :	私 は 今晩 カーディフ を 発ち ます
Baseline(B) :	i cardiff leaving tonight
Proposed(A) :	i'm leaving cardiff tonight
Source :	カーディフ は ここ から いく つめ の 停留所 です か
Baseline(B) :	how many stops is it from here cardiff
Proposed(A) :	how many stops to cardiff from here
Source :	この 列車 は カーディフ 行き です か
Baseline(B) :	this train going cardiff
Proposed(A) :	is this train bound for cardiff
Source :	カーディフ は ファッション の 中心地 です
Baseline(C) :	the center of cardiff fashion
Proposed(A) :	cardiff is the center for fashion
Source :	カーディフ 行き の 最終 は 何時 です か
Baseline(B) :	what time is the last cardiff go
Proposed(A) :	what time is the last train to cardiff

Table 6: examples both baseline and proposed method(place name test)

Source :	こんにちは 部屋 を 予約 し た コイズミ です
Baseline(C) :	hello i made a reservation for a room is koizumi
Proposed(A) :	hello this is miss koizumi . i made a reservation for a room
Source :	もしもし コイズミ 先生 は いらっしゃい ます か
Baseline(B) :	hello . is there a doctor koizumi
Proposed(A) :	hello . may i speak to ms koizumi
Source :	八 零 二 号室 の コイズミ です
Baseline(C) :	i koizumi for room eight o two
Proposed(A) :	this is ms koizumi in room eight o two
Source :	コイズミ と いい ます 。 予約 し て あり ます
Baseline(C) :	and i koizumi . i have a reservation
Proposed(A) :	my name is koizumi . i have a reservation
Source :	コイズミ 様 私 ども の ホテル へ ようこそ
Baseline(B) :	koizumi . welcome to our hotel
Proposed(A) :	mister koizumi welcome to our hotel

Table 7: Examples both baseline and proposed method(person name test)