Modality-Preserving Phrase-Based Statistical Machine Translation

Masamichi Ideue, Kazuhide Yamamoto
Department of Electrical Engineering
Nagaoka University of Technology
Nagaoka, Japan
{ideue,yamamoto}@jnlp.org

Masao Utiyama, Eiichiro Sumita
Universal Communication Research Institute
NICT
Soraku, Japan
{mutiyama,eiichiro.sumita}@nict.go.jp

***Abstract*—In machine translation (MT), modality errors are often critical. We propose a phrase-based statistical MT method that preserves the modality of input sentences. The method introduces a feature function that counts the number of phrases in a sentence that are characteristic words for modalities. This simple method increases the number of translations that have the same modality as the input sentences.**

***Keywords*—statistical machine translation; modality;**

I. Introduction

Outputs of statistical machine translation (SMT) usually contain word selection and word order errors. These errors are often evident to users because the outputs are awkward. However, there are instances in which users are likely to overlook modality errors: for example, “I do not like apples” may be translated as “I like apples.” In this case, the users would not be able to detect an error. Thus, preserving the modality of input sentences is essential.

We propose a method for preserving the modality of input sentences in phrase-based SMT (PBSMT). We introduce a special feature function in the PBSMT model that counts the number of characteristic words for modalities. Although we apply our method to Japanese-English PBSMT, it is applicable to any language pair. Our method increases the number of translations that have the same modality as the input sentences.

II. Related Studies

Previous studies in this regard have aimed to preserve sentence modality in SMT.

Finch et al. [1] divided their training data into question sentences and other sentences. They trained two models for each of the two types of the sentences. They also trained a third model with the entire training data. A class-dependent mixture was utilized to use these three models.

Goh et al. [2] proposed the use of various global features for discriminative reranking in an SMT framework. Their framework employs an online large-margin-based training algorithm for structural output support vector machines based on Margin Infused Relaxed Algorithm. The framework uses the probabilities of sentence types, such as affirmations, negations, questions, and predication.

Both of the above-mentioned studies improved the translation accuracy; however, neither of the studies discusses what expressions influence modalities. Our study focuses on characteristic modality words in negations, affirmations, and questions.

III. Translation Model Using Characteristic Modality Words

First, we manually extracted characteristic words for each modality, which are shown in Table I . We preserved the case distinctions in these words because capitalized words usually occur at the beginning of question sentences. We used the Moses toolkit [3] to tokenize English sentences.

TABLE I
Manually extracted characteristic modality words for English.

negation	not	No	cannot	’t
question	?	Why	Will	What
	Could	Is	How	Does
	Can	Do	Are	Which
	When	Where	Have	Does
	Did	Was	May	Shall

A. Translation Model

The output $\hat{e}$ of a translation is determined by the following log-linear equation in a well-known phrase-based translation model.

$$\hat{\boldsymbol{e}} = \arg\max_{\boldsymbol{e}} \sum_{i=0} \lambda_i h_i(\boldsymbol{e}, \boldsymbol{f}, \boldsymbol{c}) \tag{1}$$

where $\boldsymbol{f}$ and $\boldsymbol{e}$ are Japanese input and translation hypotheses respectively, $\boldsymbol{c}$ is the alignment between the phrases in the two languages; $h_i(\boldsymbol{e}, \boldsymbol{f}, \boldsymbol{c})$ is a feature function such as the language model and the phrase translation probability; and λ is the weight assigned to each feature function.

B. Characteristic Words in the Two Languages

The feature function defined in equation (2) only considers the output phrases. However, to preserve the modality of the input sentence, we need to consider both input and output phrases. For this purpose, we use another feature function $h(\boldsymbol{e}, \boldsymbol{f})$:

$$h(\boldsymbol{e}, \boldsymbol{f}) = \sum_i f_c(\bar{e}_i, \bar{f}_i) \tag{2}$$

$$f_c(\bar{e}, \bar{f}) = \begin{cases} 1 & if\ |\boldsymbol{C}_{\bar{e}} \cap \boldsymbol{C_E}|\&|\boldsymbol{C}_{\bar{f}} \cap \boldsymbol{C_F}| \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\bar{f}_i$ is a phrase that occurs in $\boldsymbol{f}$. Setting $\boldsymbol{C}_F$ as the set of characteristic Japanese modality words and $\boldsymbol{C}_E$ as the set of characteristic English modality words, $f_c(\bar{e}, \bar{f})$ is 1 if both $\bar{e}$ and $\bar{f}$ contain characteristic modality words for their respective languages.

The modality characteristic words in Japanese were extracted manually as shown in Table II . Japanese sentences are divided into morphemes using the morphological analyzer ChaSen[1].

TABLE II
MANUALLY EXTRACTED CHARACTERISTIC MODALITY WORDS FOR JAPANESE.

negation	nai	mase n
question	?	ka .

C. Automatic Extraction of Characteristic Words using the Log-Likelihood Ratio

The characteristic modality words shown in Tables I and II were manually extracted.

We also used a statistical measure to automatically extract characteristic words for the modalities. We used the log-likelihood ratio (LLR) [4] because the parallel corpus we chose for our experiments consisted of sentences from the travel domain; the LLR has been found to be effective for extracting characteristic words in this domain [5].

To calculate the LLR scores, we constructed a contingency table, as shown in Table III

TABLE III
CONTINGENCY TABLE USED IN THE EXTRACTION OF CHARACTERISTIC WORDS FOR THE NEGATION MODALITY.

	negation	**affirmation**
$w = 1$	a	b
$w = 0$	c	d

In Table III , w denotes the word for which an LLR score will be calculated. The value $w = 1$ signifies that w occurred in the phrase. a is the number of occurrences of w in negative sentences. while b is the number of occurrences of w in affirmative sentences. On the other hand, c is the difference of the total number of negative sentences and a, while d is the difference of the total number of affirmative sentences and b. Setting, $n = a + b + c + d$, the LLR score is defined in equation (4).

$$LLR = sign(ad - bc)LLR_0 \tag{4}$$

$$sign(z) = \begin{cases} +1 & if\ z \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

$$\begin{aligned} LLR_0 &= \frac{Pr(D|H_{dep})}{Pr(D|H_{indep})} \\ &= a \log \frac{an}{(a+b)(a+c)} + b \log \frac{bn}{(a+b)(b+d)} + \\ & c \log \frac{cn}{(a+b)(a+c)} + d \log \frac{dn}{(c+d)(b+d)} \end{aligned} \tag{6}$$

where D is the state of the parameters in Table III . In equation (6), $Pr(D|H_{indep})$ is the probability of observing the contingency table under the null hypothesis that the occurrences of the word w in the negative and affirmative sentences are independent of one another, while H_{dep} is the case in which the occurrences are dependent. The words are ranked in the decreasing order of their LLR scores.

We extracted the top N words as the characteristic words for the negation modality. The characteristic words for the question were extracted by the same method.

IV. EXPERIMENTS

A. Identification of Sentence Modality

Three modalities of parallel sentences in the corpus, *negation*, *question*, and *affirmation* were identified through the modalities of the English sentences. The identification rules were as follows:

- Negation: The sentence includes negation words shown in Table I
- Question: The sentence ends with a question mark.
- Affirmation: All other cases.

In general, the identification of the modalities of Japanese sentences is more difficult than that of English sentences. Hence, we did not use the Japanese sentences to identify the modalities of the parallel sentences.

B. Characteristic Words Extracted by LLR

Using LLR, 30 characteristic modality words were automatically extracted (LLR30). We also experimented with 20, 50, and 100 (referred to as LLR20, LLR50, and LLR100, respectively) characteristic modality words and found that the Bilingual Evaluation Understudy (BLEU) scores evaluated using these values were similar to those evaluated using 30 words. The BLEU scores are shown in Table IV , and the characteristic words extracted by LLR30 are shown in Table V .

TABLE IV
PRELIMINARY EXPERIMENTAL EVALUATION TO DECIDE THE THRESHOLD RANK FOR LLR.

	LLR20	**LLR30**	**LLR50**	**LLR100**
BLEU	32.58	32.75	32.57	32.61

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

TABLE V
CHARACTERISTIC WORDS EXTRACTED USING LLR.

Negation		**Question**	
English	Japanese	English	Japanese
’t	mase	?	ka
not	nai	you	doko
don	n	What	nani
Don	ha	How	?
didn	naka	Do	dou
can	amari	Is	ikura
doesn	mada	Can	ha
isn	ari	Where	dono
won	deki	Could	itadake
haven	ja	May	nanzi
yet	iie	do	ari
any	sonnani	Would	morae
but	wakara	’t	desyo
couldn	sonna	Are	ikaga
know	taku	does	masu
worry	koto	Will	donna
No	rare	any	kurai
wasn	naku	Which	dotira
cannot	desi	Why	yorosii
I	sinpai	there	desu
hasn	wakara	have	dore
shouldn	wakari	this	u
aren	ga	Who	itu
wouldn	mo	When	osie
anything	yoku	Does	mase
it	de	Did	ii
so	siri	don	kakari
afraid	wake	long	kono
understand	na	Shall	o
what	zannen	it	dousite

C. SMT Experiments

We used the Moses toolkit [3] for the SMT experiments. The weights of the feature functions were determined through the minimum error rate training [6].

We used the Basic Travel Expression Corpus (BTEC) English-Japanese corpus[7]. The training set consisted of 70,000 sentence pairs, and the test set included 500 sentence pairs for each modality. The development set contained 500 sentence pairs for each modality.

We compared three settings. The baseline setting used the default setting of the Moses toolkit. The MAN_E setting used the manually extracted characteristic English words used to evaluate equation (2), while MAN_EJ used the manually extracted characteristic English and Japanese words used to evaluate equation (4). LLR30 used the automatically extracted characteristic English and Japanese words to evaluate equation (4). We did not use automatically extracted words to evaluate equation (2) because the translation accuracy of MAN_EJ was better than that of MAN_E.

TABLE VI
NUMBERS OF MANUALLY EVALUATED SENTENCES WITH EACH SCORE.

	S	A	B	C	D
Baseline	60	57	34	26	93
MAN_E	66	40	38	29	97
MAN_EJ	55	54	44	29	88
LLR30	60	56	38	28	88

TABLE VII
TRANSLATION ACCURACY FOR EACH MODALITY (“AFF”, “NEG” AND “QUE” ARE AFFIRMATION, NEGATION AND QUESTION, RESPECTIVELY).

	Aff (135)	Neg (51)	Que (84)
Baseline	86.67	39.22	90.48
MAN_E	71.11	80.39	95.24
MAN_EJ	87.41	64.71	90.48
LLR30	87.41	62.75	95.24

D. Manual evaluation

We randomly extracted 90 sentence pairs to test the methods: none of these occurred in the evaluation set for the manual evaluation. The result of the evaluation is shown in Table VI and Table VII . The column labels columns “S”, “A”, “B”, “C” and “D” in Table Table VI indicate the grades for the translation accuracy. The respective grades are defined as below:

- S: Completely perfect and sounds like it was spoken by a native speaker.
- A: While completely grammatically correct, it does not sound like it was spoken by a native speaker.
- B: There are grammatical mistakes, but it is easy to understand and contains all of the information from the original sentence.
- C: Contains many grammatical mistakes and is very difficult to understapppnd.
- D: Either it is completely mistranslated or it is completely incomprehensible.

Table VII shows the percentage of the outputs that preserved the modality of the input; the numbers in parentheses column headings indicate the number of sentence-pairs for each modality.

As seen in Table VI , all the methods have the same translation accuracy if we assume that the output grades “S”, “A” and “B” indicate good translations.

E. Accuracy of each modality

As indicated by Table VII , the accuracy of the negation modality for the baseline is 39.22%, while those of the other three methods indicate a marked improvement. An example of the improved output achieved by our proposed method is shown below.

- Input: sa-kasu to doubutu en, dotti ni iko u ka . (which means “The circus or the zoo, which shall we go to?”)
- Baseline: Let’s go to the circus and, the zoo? (×)
- MAN_EJ: Which one shall we go to the circus and Zoo?

In this example, the input has the question modality, but the output produced by the baseline is not a question. On the other hand, the output produced by MAN_EJ has the characteristic question word “Which”. In other words, it has the question modality. As our proposed method of considering the characteristic words for each modality tends to select phrases that include characteristic words, the output tend to include the characteristic words, and thus preserves the modality of the input.

In this case of affirmative sentences, the translation accuracy of MAN_E was lower than that of the others. We only added the features ϕ_{neg} and ϕ_{que} to equation (2), with no such consideration for affirmative sentences; hence, our proposed method tends to select phrases that include characteristic words for negations or questions if the weight of ϕ_{neg} or $\phi_{characteristic}$ is non-negative. MAN_E, which considers only characteristic words of English, incorrectly selects phrase pairs in which the English phrase contains the characteristic words. On the other hand, MAN_EJ is constrained by the requirement that both the Japanese and English phrases in the phrase pair include the characteristic word, which prevents incorrect selection.

The accuracy of LLR30 was better than the baseline accuracy of all modalities. Even when the characteristic words are automatically extracted, our proposed method can improve the accuracy of preserving sentence modality in a translation.

F. Examples of Incorrect Translations

In this section, we discuss examples of incorrect translations produced by our proposed method. The following is an example in which the output of MAN_E is a question because the tag question "isn't it?" has been added.

- Input: yasasiku utte kudasai ne . (Please go easy.)
- MAN_E: Please go easy, isn't it? (×)
- MAN_EJ: Please go easy.

The expression "isn't it?" includes a characteristic question word "?" and a characteristic negation word "'t", and therefore, a tag question tends to be in the translation. MAN_EJ which both the Japanese and English words in a phrase pair, produced a correct translation.

The following is an example in which the translation produced by MAN_EJ has the negation modality despite the question modality of the input:

- Input: kyanseru si te mo kamai mase n ka . (May I cancel it?)
- Baseline: May I cancel?
- MAN_EJ: I don't mind if you cancel it? (×)

The reason for this is that the input includes the characteristic negative modality words "mase n" and the question word "ka .", "mase n" was manually extracted as a phrase with a strong negative characteristic. However, as this example show, the phrase does not always express this modality.

V. Conclusion

In this paper, we proposed a method that adds a feature function to the phrase-based statistical translation model that identifies characteristic words for modalities, and we compared our method with a baseline using both manually and automatically extracted words.

Our experimental results demonstrate that our method produces more translations that retain the modality of the input sentence than the baseline does. In particular, our method improved the translation of sentences with the negation modality.

We also compared the results obtained by the two methods for extracting characteristic words: manual extraction and automatic extraction using LLR. We confirmed that automatic extraction performed the same as or better than manual extraction even though the automatically extracted characteristic words included noise.

References

[1] A. Finch, E. Sumita, and S. Nakamura, "Class-Dependent Modeling for Dialog Translation," *IEICE TRANSACTION on Information and Systems*, vol. E92-D, no. 12, pp. 2469-2477, 2009.

[2] C. Goh, T. Watanabe, A. Finch, and E. Sumita, "Discriminative Reranking for SMT using Various Global Features," In *Proceedings of 4th International Universal Communication Symposium (IUCS 2010)*, 2010, pp.8–14.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL)*, 2007, pp. 177–180.

[4] D. Ted, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74.

[5] K. Chujo, M. Utiyama, and K. Oghigian, "Selecting Level-Specific Kyoto Tourism Vocabulary Using Statistical Measures," *New Aspects of English Language Teaching and Learning*, pp. 126–138.

[6] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics(ACL)*, 2003, pp. 160–167.

[7] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World." In *Proceedings of International Conference on Language Resources and Evaluation*, 2002, pp. 147–152.