FUN-NRC: Paraphrase-augmented Phrase-based SMT Systems for NTCIR-10 PatentMT

Atsushi Fujita*
Future University Hakodate (FUN)
116-2, Kameda-nakano-cho
Hakodate, Hokkaido 041-8655, Japan
fujita@fun.ac.jp

Marine Carpuat
Information and Communication Technologies
National Research Council (NRC)
Ottawa, Ontario K1A 0R6 Canada
marine.carpuat@nrc.gc.ca

ABSTRACT

This paper describes FUN-NRC group's machine translation systems that participated in the NTCIR-10 PatentMT task. The central motivation of this participation was to clarify the potential of automatically compiled collections of sub-sentential paraphrases. Our systems were built using our baseline phrase-based SMT system by augmenting its phrase table with novel translation pairs generated by combining paraphrases with translation pairs learned directly from the training bilingual data. We investigated two methods for phrase table augmentation: source-side augmentation and target-side augmentation. Among the systems we submitted, the two that worked best were (a) the one that paraphrased only unseen phrases into translatable phrases at the source side and (b) the one that paraphrased target phrases only into phrases that were seen in the original phrase table. Both these systems were trained on not only bilingual, but also monolingual data. The other two systems were trained using only bilingual data. This paper also reports on our follow-up experiments focusing on the relationship between reordering restriction and system performance.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Machine Translation

General Terms

Experimentation, Performance

Keywords

Machine Translation, Phrase-based SMT, Paraphrase

Team Name

[FUN-NRC]

Subtasks/Languages

[Japanese-to-English][English-to-Japanese]

External Resources Used

[MeCab][SRILM][MGIZA++][Vowpal Wabbit]

1. INTRODUCTION

Given a large-scale bilingual parallel corpus in a particular domain, there are two promising approaches to build a machine translation (MT) system for that domain without making an enormous effort to tailor an in-domain system with domain-specific translation dictionaries and translation rules [22]. One is to build a statistical MT (SMT) system on the bilingual corpus and the other is to create a statistical post-editor for a particular rule-based MT (RBMT) system. Both of these two have shown good results in the past NTCIR PatentMT evaluation [17].

In the SMT community, incorporating higher levels of linguistic knowledge, such as those about syntax, semantic, and discourse, has been recognized as an important issue. Structure-aware SMT systems, including syntax- and dependency-based ones, that benefit from mature analyzers, have achieved promising results in the past and current NTCIR PatentMT evaluations [17, 18].

On the other hand, there is another important issue: how to create a translation system from textual data relying only on minimal knowledge about language and domain of interest. This is crucial for particular situations, including translating from/to resource-poor languages and applying SMT to a new domain.

Among several research directions, in the NTCIR-10 PatentMT task, we focused on exploiting automatically compiled semantic resources, i.e., collections of sub-sentential paraphrases, with a phrase-based SMT framework. The collections of paraphrases that we investigated were created using both of the provided in-domain bilingual and monolingual corpora. Our approach was quite domain-independent and language-independent. No domain specific knowledge was introduced. The only linguistic information used for building the systems was that contained within the preprocessing tools and stop word lists used for sanitizing paraphrase collections.

2. BASE SYSTEM

2.1 Preprocessing for the Provided Corpora

In the NTCIR-10 PatentMT task, sentence-aligned bilingual data and unprocessed monolingual data were distributed by the organizers to the participants. We extracted textual data from the provided monolingual data, and we then extracted sentences from the textual data. Sentences in the Japanese data were heuristically extracted, while sentence splitting for English data was performed by NRC's in-house tool. Finally, sentences in both monolingual and bilingual corpora were tokenized using MeCab[1] for Japanese sentences and NRC's in-house tokenizer for English sentences.

2.2 Top-level Architecture

FUN-NRC systems were built using NRC's in-house phrase-based SMT system, "PortageII 1.0." This system contains no explicit linguistic knowledge.

Decoding of log-linear combination of models was performed by cube-pruning [21] with a predetermined distortion limit of 7 words.

*During the development and evaluation phase of the shared task, the first author stayed at NRC as a JSPS (the Japan Society for the Promotion of Science) Postdoctoral Fellow for Research Abroad.

[1] http://mecab.googlecode.com/, version 0.994 with IPAdic

The weights of the component models were optimized using lattice-based batch version of MIRA [8], with BLEU score [31] against the provided development data `pat-dev-2006-2007.txt` used as the objective function. Calculation of BLEU scores is performed on the raw system output case-insensitively without re-tokenizing sentences as in the official evaluation.

None of the hyper parameters of both the top-level system and component models had been tuned for this shared task[2]. They were all determined during NRC's previous participation in Chinese-to-English and Arabic-to-English translation tasks at NIST OpenMT 2012[3].

2.3 Language Models

Two 5-gram language models were built with Kneser-Ney smoothing using SRILM[4]. One was made from the target side of the bilingual corpus, while the other was trained on the monolingual corpus of the target language. The two models were log-linearly combined at the top level of the system.

2.4 Phrase Table

Three word alignment algorithms were employed. While IBM Model 2 (IBM2) [4] and HMM [20] alignments were determined by our implementation, IBM Model 4 (IBM4) [4] alignments were obtained using MGIZA++[5].

Phrase alignments were identified from each of the word-aligned bilingual corpus by the heuristic "grow-diag-final" [24] with a maximum phrase length of 8. Joint counts of each phrase-level translation pair derived from the above three alignment algorithms were summed up with a binary indicator feature for each source algorithm added (see also Section 3.4).

Finally, for each translation pair, the following two types of conditional phrase probabilities for both forward and backward directions were calculated.

- Kneser-Ney estimates using the joint counts [7]
- Lexical estimates based on HMM word alignment [34]

2.5 Distortion Models

In addition to the standard distance-based distortion feature, the following two models were incorporated into the log-linear model.

- Lexicalized distortion model [25]
- Hierarchical lexicalized distortion model [16] with an unrestricted shift-reduce parser [9]

The probability of each orientation (monotone, swap, and discontinuous) was first estimated for each of word alignment algorithms, IBM2, HMM, and IBM4, respectively. The three results were then combined through the "fill-up" strategy [2] to compile the final models.

2.6 True-casing

We call the process of restoring proper mixed case to the MT output "true-casing." This process was performed by applying the decoder for MT hypothesis generation to the following two component models both of which were built using the target-side of the bilingual corpus.

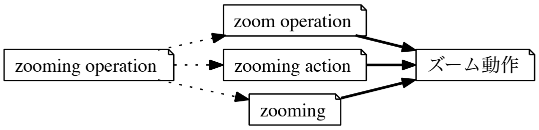

Figure 1: Source-side phrase table augmentation: paraphrases of the source language (English in this case) are connected to source phrases of the phrase table.

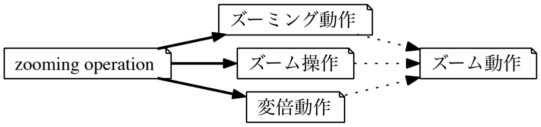

Figure 2: Target-side phrase table augmentation: paraphrases of the target language (Japanese in this case) are connected to target phrases of the phrase table.

- Word-level translation (case-mapping) probabilities between true-case and normalized case
- True-case 3-gram language model

Unlike the decoding of translation hypothesis, weights for this process were manually predetermined.

3. PARAPHRASE-AUGMENTED SYSTEMS

Incorporating higher levels of linguistic knowledge has been recognized as one of the most challenging issues in the SMT community. To address this issue from a rather language-independent and domain-independent viewpoint, in this exercise, we concentrated on exploiting automatically compiled collections of sub-sentential paraphrases.

Among several conceivable approaches (see Section 7 for detail), we used paraphrases in one language, in combination with translation pairs that had been directly obtained from bilingual corpus (henceforth, **original translation pairs**), to generate novel **fabricated translation pairs** as in [5, 28]. Figure 1 illustrates a fabricated translation pair obtained by augmenting the source side of the phrase table, where solid arrows indicate original translation pairs, while dotted arrows indicate paraphrases in the source language. A fabricated translation pair ("zooming operation", "ズーム動作") is obtained by combining a paraphrase pair in the source language, e.g., ("zooming operation", "zoom operation") and an original translation pair, e.g., ("zoom operation", "ズーム動作"). In what follows, we refer to the phrase that connects fabricated translation pairs, e.g., "zoom operation" in the example just given, as the **proxy phrase**. Fabricated translation pairs can also be obtained by using paraphrases in the target language as shown in Figure 2.

The fabricated translation pairs were then used to augment the phrase table of our base system. In the development phase, we investigated how to incorporate the fabricated translation pairs along with the original ones. The rest of this section discusses the following four topics related to this investigation.

1. Paraphrase collections
2. Score of individual paraphrase pair
3. Aggregation of multiple paths
4. Paraphrase-related features for decoding

[2] After the formal run was closed, we made an investigation into distortion limit. See Section 6.2 for the details.

[3] http://www.nist.gov/itl/iad/mig/openmt12.cfm

[4] http://www.speech.sri.com/projects/srilm/, version 1.5.7

[5] http://sourceforge.net/projects/mgizapp/, version 0.6.3

3.1 Paraphrase Collections

We investigated the following three collections of automatically acquired sub-sentential paraphrases.

P_{Seed}: Paraphrases acquired from bilingual corpora relying on the assumption that expressions in one language that share translations in the other side of the language are semantically equivalent [1]. For instance, see the middle and right nodes in Figure 1. A pair of phrases ("zoom operation", "zooming action") is extracted as paraphrases because they share a translation "ズーム動作." The initial collection obtained by this principle was then cleaned by using stop word lists of both languages and several heuristics to filter out unreliable pairs [15].

P_{Hvst}: Paraphrases acquired using paraphrase patterns derived from P_{Seed} [15]. For instance, given a paraphrase pair ("zooming operation", "zooming action"), a pattern "X operation" $\Leftrightarrow$ "X action" was induced. Then, other words that match the variable X of both sides of the pattern were extracted from the monolingual corpus. For the above pattern, "programmed" and "regenerative" were obtained. Finally, new paraphrase pairs ("programmed operation", "programmed action") and ("regenerative operation", "regenerative action") were generated.

P_{OOPH}: This collection was created from P_{Hvst}. Consider the direction of paraphrases $ph_1 \Rightarrow ph_2$. For source-side phrase table augmentation, only pairs such that ph_1 had never appeared in the source side of phrase table were retained. Similarly, for target-side phrase table augmentation, only pairs such that ph_2 had never appeared in the target side of phrase table were retained.

One of our contributions is that we have used the in-domain monolingual corpus as a source of paraphrases, while it has been typically used only for creating language models.

3.2 Score of Individual Paraphrase Pair

The reliability of each paraphrase pair, i.e., $Para(ph_1 \Rightarrow ph_2)$, can be measured in several ways, including the paraphrase probability based on shared translations [1] and the contextual similarity computed on a monolingual corpus [19].

In the initial work on augmenting phrase table with paraphrases [5], the paraphrase probability calculated based on shared translation (henceforth, ***PivProb***) was employed to let the decoder know the quality of the paraphrase pair. However, this score can be estimated only for P_{Seed}, but not for P_{Hvst} and P_{OOPH}. In addition, this estimate may not be accurate because of the limited size of bilingual corpora.

On the other hand, contextual similarity would be more accurate because there is far more monolingual than bilingual data available. Furthermore, this score can be computed for any pair of phrases that appear in a given monolingual corpus. Marton et al. [28] employed contextual similarity for scoring paraphrase pairs acquired from a monolingual corpus, and alleviated the out-of-vocabulary problem at the source side of translation in the SMT framework.

Although numerous measures for contextual similarity (features, weighting and filtering schemes, and aggregation formulae) have been proposed, the best measure would be different depending on the task, language, and domain. In our exercise, simply following [15], we employed cosine measure between context vectors comprising the counts of adjacent 1- to 4-grams of each token within the monolingual corpus and the corresponding side of the bilingual corpus (henceforth, ***CosSim***).

We compared *PivProb* and *CosSim* in the development phase (see Section 4). We also used both *PivProb* and *CosSim* for filtering paraphrase collections: paraphrase pairs with low scores were removed. For the threshold value on *PivProb* (th_p), we conformed to the convention, i.e., 0.01 [13, 29, 12]. For the threshold value on *CosSim* (th_s), on the other hand, we used an arbitrary value 0.1.

3.3 Aggregation of Multiple Paths

In previous work, there is no description of how to estimate translation probabilities for the fabricated translation pairs, nor how to encode the fabricated translation pairs generated relying on more than one proxy phrase such as those illustrated in Figures 1 and 2.

In our systems, translation probabilities for the fabricated translation pairs were given by aggregating the information of all involved proxy phrases as follows.

Source-side fabricated pairs: Let s' be the source-side phrase of the fabricated translation pair and S be the set of proxy phrases at the source side. In the case of Figure 1, for instance, s' is "zooming operation" in the left and S is the set of the three English phrases in the middle.

$$p(t|s') = \frac{\sum_{s \in S} \Big(p(t|s) Para(s' \Rightarrow s) \Big)}{\sum_{s \in S} Para(s' \Rightarrow s)}, \quad (1)$$

$$p(s'|t) = \frac{\sum_{s \in S} \Big(p(s|t) Para(s \Rightarrow s') \Big)}{\sum_{s \in S} Para(s \Rightarrow s')}. \quad (2)$$

Target-side fabricated pairs: Let t' be the target-side phrase of the fabricated translation pair and T be the set of proxy phrases at the target side. In the case of Figure 2, for instance, t' is "ズーム動作" in the right and T is the set of the three Japanese phrases in the middle.

$$p(t'|s) = \frac{\sum_{t \in T} \Big(p(t|s) Para(t \Rightarrow t') \Big)}{\sum_{t \in T} Para(t \Rightarrow t')}, \quad (3)$$

$$p(s|t') = \frac{\sum_{t \in T} \Big(p(s|t) Para(t' \Rightarrow t) \Big)}{\sum_{t \in T} Para(t' \Rightarrow t)}. \quad (4)$$

The above estimates are no longer probabilities; for instance, $\sum_t p(t|s')$ is not guaranteed to be 1. We thus call them **translation scores**. In contrast, lexical estimates [34] for both directions were also calculated in the same manner for the original translation pairs.

The combination of paraphrases and original translation pairs also generates other original translation pairs. Although exploiting them to give a bonus would be an interesting topic, we ignored them for now.

3.4 Paraphrase-related Features for Decoding

The additional features were two-fold: binary features for a classification of translation pairs, and paraphrase scores.

3.4.1 Translation Pair Indicators

To explicitly tell the decoder where a given translation pair came from, the following binary features were assigned.

- Source of the translation pair:
  - Original translation pair obtained by a particular word alignment algorithm (IBM2, HMM, and IBM4)
  - Fabricated translation pair generated using paraphrase pairs in a particular collection (P_{Seed} and P_{Hvst} (or P_{OOPH}))

Table 1: Feature encoding scheme for translation pairs.
$^{(*1)}$Kneser-Ney smoothed probability based on joint counts.
$^{(*2)}$At least one of three features is True. $^{(*3)}$At least one of two features is True.

Feature	Translation pair type		
	Original	Source-side fabricated	Target-side fabricated
(a1) Forward translation score	KN$^{(*1)}$	See Eq. (1)	See Eq. (3)
(a2) Backward translation score	KN$^{(*1)}$	See Eq. (2)	See Eq. (4)
(b1) IBM2 alignment oriented	True/False$^{(*2)}$	False	False
(b2) HMM alignment oriented	True/False$^{(*2)}$	False	False
(b3) IBM4 alignment oriented	True/False$^{(*2)}$	False	False
(c1) Fabricated using P_{Seed}	False	True/False$^{(*3)}$	True/False$^{(*3)}$
(c2) Fabricated using P_{Hvst} or P_{OOPH}	False	True/False$^{(*3)}$	True/False$^{(*3)}$
(d1) Unseen in the phrase table	False	True/False	True/False
(d2) Unseen in the bilingual corpus	False	True/False	True/False
(e1) Forward paraphrase score for source-side augmentation	1	See Section 3.4.2	1
(e2) Backward paraphrase score for source-side augmentation	1	See Section 3.4.2	1
(e3) Forward paraphrase score for target-side augmentation	1	1	See Section 3.4.2
(e4) Backward paraphrase score for target-side augmentation	1	1	See Section 3.4.2

- Whether the fabricated phrase (s' or t') is registered in the original phrase table or not
- Whether the fabricated phrase (s' or t') is seen in the bilingual corpus or not

3.4.2 Paraphrase Scores

The following four scores were introduced to inform the decoder about the reliability of the paraphrase pairs used to generate the fabricated translation pair. The value of each feature is a real number within $[0, 1]$.

- Source-side forward score: $\frac{1}{|S|}\sum_{s\in S} Para(s' \Rightarrow s)$
- Source-side backward score: $\frac{1}{|S|}\sum_{s\in S} Para(s \Rightarrow s')$
- Target-side forward score: $\frac{1}{|T|}\sum_{t\in T} Para(t \Rightarrow t')$
- Target-side backward score: $\frac{1}{|T|}\sum_{t\in T} Para(t' \Rightarrow t)$

In case a symmetric metric, such as *CosSim*, is used for the score of individual paraphrase pair, forward and backward scores in the above formulae are identical; so they can be de-duplicated.

3.5 Summary of features

Table 1 summarizes how the features of three different (exclusive) types of translation pairs were encoded in the phrase table. The notions of the translation probabilities for original translation pairs and translation scores for the fabricated translation pairs were different. However, they were exclusively fired in combination with binary indicator features.

4. DEVELOPMENT

To determine the systems to be submitted, we conducted experiments using held-out data from the past two NTCIR PatentMT evaluations, i.e., `ntc7-fmlrun` and `ntc8-fmlrun`. We used BLEU score for the performance measure.

Table 2 summarizes the number of paraphrase pairs in our paraphrase collections P_{Seed} and P_{Hvst}. As described in Section 3.2, only reliable pairs, which were determined on the basis of *PivProb* and *CosSim*, were used for augmenting the given phrase table. Note that P_{Hvst} was extremely larger than P_{Seed}. Thus, in our experiment, we restricted the paraphrase pairs to those that could potentially be used for translating the given held-out and test data in a way that caused the given phrase table to be augmented, and computed *CosSim* only for them.

Table 2: # of paraphrase pairs acquired using both bilingual and monolingual corpora (refer to [15] for detail).

	Threshold		English	Japanese
	th_p	th_s		
P_{Seed}	0	0	7,154,449	5,142,634
P_{Seed}	0.01	0.1	1,136,765	756,434
P_{Hvst}	0.01	0	272,388,773	142,526,447
P_{Hvst}	0.01	0.1	n/a	n/a

Table 3: Average BLEU scores (raw output; case-insensitive) over two held-out data: `ntc7-fmlrun` and `ntc8-fmlrun`.

System	$Para(\cdot \Rightarrow \cdot)$	Ja-to-En	En-to-Ja
Base system	-	33.30	37.64
Saug-P_{Seed}	*PivProb*	33.65	37.98
Saug-P_{Seed}	*CosSim*	33.27	37.73
Saug-P_{Hvst}	*CosSim*	33.22	37.89
Saug-P_{OOPH}	*CosSim*	**33.72**	**38.16**
Saug-P_{Seed}+P_{Hvst}	*CosSim*	32.91	37.76
Taug-P_{Seed}	*PivProb*	33.34	37.64
Taug-P_{Seed}	*CosSim*	**33.56**	**38.19**
Taug-P_{Hvst}	*CosSim*	33.43	37.98
Taug-P_{OOPH}	*CosSim*	33.21	38.08
Taug-P_{Seed}+P_{Hvst}	*CosSim*	32.99	37.53

Table 3 shows the results for several systems, where “Saug” and “Taug” indicate the source-side augmented and target-side augmented systems, respectively. As the score of individual paraphrase pairs in P_{Seed}, *PivProb* was better than *CosSim* when it was used at the source side. On the other hand, at the target side, *CosSim* yielded better results than *PivProb*.

Among the systems we developed, for the source-side augmentation, Saug-P_{OOPH} achieved the highest BLEU score. This means

Table 4: # of paraphrase pairs acquired using only bilingual corpus.

	Threshold		English	Japanese
	th_p	th_s		
P_{Seed}	0	0	7,154,449	5,142,634
P_{Seed}	0.01	0.1	910,693	545,924
P_{Hvst}	0.01	0	10,182,504	5,382,207
P_{Hvst}	0.01	0.1	3,800,022	1,914,799

that paraphrasing only unseen phrases into translatable phrases had the largest impact. Generating phrases in the target language that were unseen in the bilingual corpus expands the search space. However, for the target-side augmentation, Taug-P_{Hvst} and Taug-P_{OOPH} did not beat Taug-P_{Seed} (with *CosSim*).

On the basis of the results, we decided to submit the following two best systems for both Japanese-to-English and English-to-Japanese tracks. As both systems use *CosSim*, it is no longer described in the rest of this paper.

Saug-P_{OOPH}: Phrase-based SMT system augmented with paraphrases. This system paraphrases only unseen phrases in the source language into translatable phrases using P_{OOPH}.

Taug-P_{Seed}: Phrase-based SMT system augmented with paraphrases. This system paraphrases target phrases into phrases that were seen in the original phrase table using P_{Seed}.

While the above two systems used both the bilingual and monolingual corpora, all participants with data-driven systems were obliged to submit at least one system developed using only the bilingual corpus. We thus also submitted the following two systems.

Const-Saug-P_{Hvst}: Phrase-based SMT system augmented with the paraphrase collection P_{Hvst}. We created the collection regarding only the source side of the bilingual corpus as a monolingual corpus. The language model was also created only from the target side of the bilingual corpus. As the bilingual corpus was significantly smaller than the monolingual corpus, P_{Hvst} available for this system was also limited as shown in Table 4 (cf. Table 2).

Const-mixLM: Phrase-based SMT system with adaptation through document-level linear mixtures of language models [6]. Mixture components consist of $(k + 1)$ target-side 5-gram language models learned on a partition of the bilingual corpus into k clusters, and on the entire target side of the bilingual corpus. Documents in the bilingual corpus were clustered using k-means on their T-dimensional topic distributions, learned using the Vowpal Wabbit[6] implementation of LDA. We used $T = 500$ and $k = 16$, which were determined using held-out data. Mixture weights were learned using the EM algorithm to fit a mixture of source-side 3-gram language models to each document to be decoded [14].

5. OFFICIAL RESULTS

5.1 Results of Intrinsic Evaluation

In the NTCIR-10 PatentMT task, human evaluation was regarded as the primary evaluation. However, it was conducted only for the selected systems. Fortunately, our primary system, Saug-P_{OOPH}, was included in both adequacy and acceptability evaluations [18] and got the results shown in Table 5. The adequacy was judged by average scores on a 5-point scale, while the acceptability score shows a win-rate in pairwise comparison between the systems based on a classification-based human judgment. Both scores were calculated based on the translations for the same 300 sentences.

Table 5: Official results of intrinsic evaluation (human).

Track	System	Adequacy	Acceptability
Ja-to-En	Saug-P_{OOPH}	2.89	0.43
En-to-Ja	Saug-P_{OOPH}	2.67	0.38

Table 6: Official results of intrinsic evaluation (automatic). $^{(*)}$Submitted results were inappropriate (see Section 6.1).

Track	System	BLEU	NIST	RIBES
Ja-to-En	Saug-P_{OOPH}	0.3156	8.2507	0.6955
	Taug-P_{Seed}	0.3165	8.2198	0.6929
	Const-Saug-P_{Hvst}	0.3058	8.1114	0.6911
	Const-mixLM	0.3065	8.1400	0.6906
En-to-Ja	Saug-P_{OOPH}	0.3422	8.2345	0.7096
	Taug-P_{Seed}	0.3405	8.2116	0.7089
	Const-Saug-P_{Hvst}	0.3289	8.0977	0.7048
	Const-mixLM$^{(*)}$	0.2259	7.1185	0.6651

Table 7: Official results of the chronological evaluation (automatic).

Track	System	BLEU	NIST	RIBES
Ja-to-En	Saug-P_{OOPH}	0.3120	8.0016	0.6970
En-to-Ja	Saug-P_{OOPH}	0.3357	8.0533	0.7087

On the adequacy evaluation, our primary system, Saug-P_{OOPH}, was ranked 10th among the selected 18 systems in the Japanese-to-English track, and 10th among the selected 14 systems in the English-to-Japanese track. On the acceptability evaluation, our primary system was ranked 8th among the selected 9 systems in both tracks. The results indicate that our submitted systems have a large room for improvement. The descriptions of superior systems suggest that incorporating syntactic information helped a lot.

Here we would raise a question: how reliable were the results of human evaluation? Evaluators' language proficiency in both of source and target languages and the level of their expertise in evaluating MT outputs are unknown. Showing the reference sentences to the evaluators might bias their decision. Evaluation methodology itself must be justified in a proper manner; nevertheless, no measure of inter-annotator agreement, e.g., Cohen's κ [11], was shown in the official results.

Table 6 summarizes the official automatic evaluation results. According to BLEU and NIST scores, our first two systems finished in the top third of all submissions, beating all of the six baseline systems. On the other hand, these two systems were ranked much lower places according to RIBES score. The results suggest that our systems were relatively good at generating phrase-level translations, but lacked the ability to deal with their ordering.

Given the high correlation between RIBES score and human evaluation score observed in the past NTCIR PatentMT evaluation [17], we envisage that improvement of word/phrase ordering is vital to obtain translations with better quality.

5.2 Results of Chronological Evaluation

Table 7 summarizes the scores of automatic evaluation metrics for the chronological evaluation data, i.e., `ntc9-fmlrun`. We

[6] http://hunch.net/~vw/, version 6.1

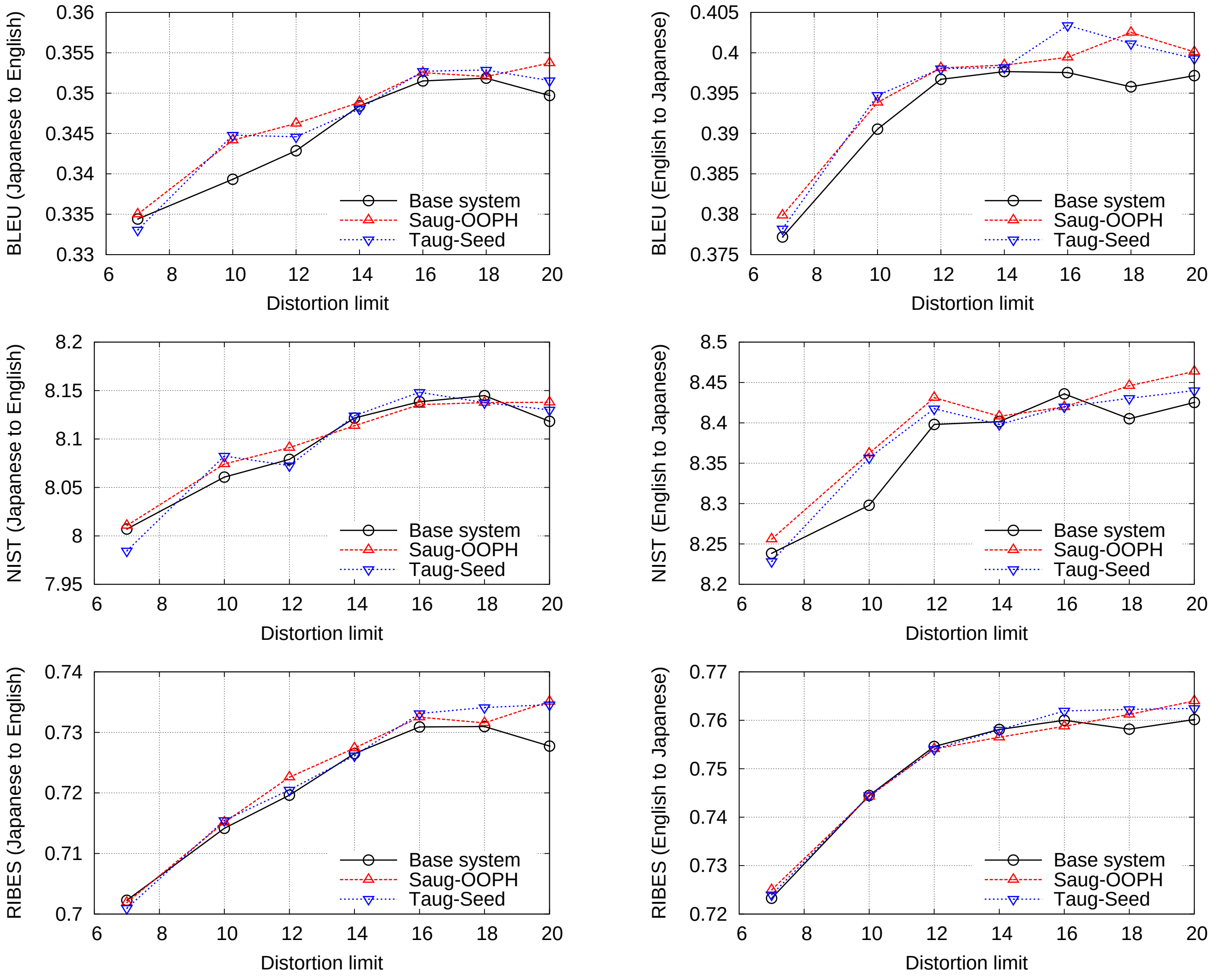

Figure 3: Averaged scores (raw output; case-insensitive) over randomized 3 runs and two held-out data: `ntc7-fmlrun` and `ntc8-fmlrun`. Three figures in the left show the results for Japanese-to-English task, while the right three are for English-to-Japanese task. The leftmost points in the BLEU score figures correspond to those in Table 3, but smoothed with multiple runs.

observed the same trend as the intrinsic evaluation results: our primary system, Saug-P_{OOPH}, was placed relatively high rank according to BLEU and NIST scores, but was ranked in the middle according to RIBES score.

6. POST-EVALUATION EXPERIMENTS

6.1 Results with a Bug-fixed System

After the formal run was closed, we found that our 4th system, Const-mixLM, for the English-to-Japanese track had generated translations with non-tuned weights. Translations generated using the tuned weights resulted in higher scores than those of the submitted system: BLEU, NIST, and RIBES scores under the official measurement procedure were 0.3303, 8.1101, and 0.7051, respectively.

6.2 Investigation into Distortion Limit

For the official submissions, we adopted predetermined values of the hyper parameters. However, the value for the distortion limit, i.e., 7, might be too restrictive to account for long-distance reordering, which often happens between far distant languages [26], i.e., Japanese and English in this case. As described in Section 5, the different trends between BLEU and RIBES scores also suggest that our systems lacked good reordering ability.

We conducted a clarification experiment with our first two systems and our base system. To improve reliability of the results, we performed weight tuning for each setting 5 times over the original and 4 randomized sub-samples of the development data [8], and averaged the results over the 5 runs for the two held-out data, i.e., `ntc7-fmlrun` and `ntc8-fmlrun`. As shown in Figure 3, by relaxing the distortion limit, our systems achieved higher scores in automatic evaluation. We used only BLEU score as the objective function of optimization same as the submitted systems. Nevertheless, NIST and RIBES scores were also dramatically improved. The results, especially a large gain in RIBES score, indicate that the long-distance reordering can at least partially be accounted for, still without any language-specific knowledge. The results also demonstrate that our paraphrase-augmented systems, depicted with red

Table 8: Post-evaluation results of intrinsic evaluation (automatic; with official setting).

Track	System	BLEU	NIST	RIBES
Ja-to-En	Base system	0.3259	8.3408	0.7151
	Saug-P_{OOPH}	0.3238	8.3169	0.7122
	Taug-P_{Seed}	0.3255	8.3470	0.7129
En-to-Ja	Base system	0.3566	8.3274	0.7392
	Saug-P_{OOPH}	0.3573	8.3799	0.7370
	Taug-P_{Seed}	0.3581	8.3485	0.7366

Table 9: Post-evaluation results of chronological evaluation (automatic; with official setting).

Track	System	BLEU	NIST	RIBES
Ja-to-En	Base system	0.3247	8.0655	0.7191
	Saug-P_{OOPH}	0.3215	8.0301	0.7134
	Taug-P_{Seed}	0.3228	8.0505	0.7160
En-to-Ja	Base system	0.3482	8.1236	0.7355
	Saug-P_{OOPH}	0.3497	8.1745	0.7339
	Taug-P_{Seed}	0.3478	8.1158	0.7356

and blue lines, have achieved better performance than our base system, even with a larger distortion limit.

We would have achieved significantly better results if we had determined the optimal value for the distortion limit of our base system first. According to the averaged BLEU scores obtained through the above investigation, we chose 18 for the Japanese-to-English task and 14 for the English-to-Japanese task. Then, unlike the investigation, we chose the single set of tuned weights that achieved the median BLEU scores among 5 runs [10]. Tables 8 and 9 show the improved automatic evaluation scores for intrinsic and chronological evaluation data, respectively (cf. Tables 6 and 7). These scores were calculated under the official measurement procedure.

7. RELATED WORK

Integration of paraphrases into MT systems has been popular as a way to improve coverage and accuracy. Previous work in this line are classified into the following five categories.

- Rewriting input sentences before putting them into a given MT system [33, 32]
- Augmenting input sentences with paraphrases [13, 30, 23]
- Enlarging the given training bilingual corpus by generating paraphrased sentences [3]
- Augmenting phrase tables to translate OOVs via their paraphrases [5, 28]
- Tuning weights of log-linear models referring to the increased number of paraphrased references [27]

Our approach to augment the source side of the phrase table is classified into the fourth. On the other hand, to the best of our knowledge, augmenting the target side of the phrase table has never been examined.

8. CONCLUSION

In the NTCIR-10 PatentMT task, we have investigated automatically compiled collections of sub-sentential paraphrases, with a standard phrase-based SMT framework. In the development phase, we revealed that our paraphrase collections improve the performance of MT systems at least in the automatic evaluation. Among several systems we developed, two systems worked best: (a) the one that paraphrased only unseen phrases into translatable phrases at the source side and (b) the one that paraphrased target phrases only into phrases that were seen in the original phrase table.

To collect human judgments, we submitted the above two systems for both Japanese-to-English and English-to-Japanese tracks, together with two other systems that were trained using only the bilingual corpus. Only one of our systems was evaluated by humans in each track, and the results showed that they were beaten by the other systems that take sentence structure into account. Relatively low RIBES scores suggest that our systems lacked good reordering ability. On the other hand, according to BLEU and NIST scores, our main two systems were ranked in the top third of all submissions, beating all of the six baseline systems. This suggests that our systems were relatively good at generating phrase-level translations.

Motivated by the above findings, we conducted post-evaluation experiments focusing on the distortion limit. This revealed that searching a wider range of hypotheses with a relaxed distortion limit makes the performance of our phrase-based SMT systems comparable to some of the other systems that take syntax information into account according to automatic measures.

9. ACKNOWLEDGMENTS

The first author is deeply grateful to Pierre Isabelle and Michel Mellinger for their efforts in making it possible for him to use NRC's PortageII for the NTCIR-10 PatentMT task. The results could not have happened without the insightful comments, suggestions, and technical support from the Portage team at National Research Council Canada: Boxing Chen, Colin Cherry, George Foster, Eric Joanis, Howard Johnson, Roland Kuhn, Samuel Larkin, and Darlene Stewart.

10. REFERENCES

[1] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 597–604, 2005.

[2] Arianna Bisazza, Nick Ruiz, and Marcello Federico. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of International Workshop of Spoken Language Translation (IWSLT)*, pp. 136–143, 2011.

[3] Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. Improving statistical machine translation by paraphrasing the training data. In *Proceedings of International Workshop of Spoken Language Translation (IWSLT)*, pp. 150–157, 2008.

[4] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[5] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 17–24, 2006.

[6] Marine Carpuat, Cyril Goutte, and Pierre Isabelle. Filtering and routing multilingual documents for translation. In *Proceedings of the 2012 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 2012.

[7] Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, 2011.

[8] Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 427–436, 2012.

[9] Colin Cherry, Robert C. Moore, and Chris Quirk. On hierarchical re-ordering and permutation parsing for phrase-based decoding. In *Proceedings of the 7th Workshop on Statistical Machine Translation (WMT)*, pp. 200–209, 2012.

[10] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 176–181, 2011.

[11] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46, 1960.

[12] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, pp. 85–91, 2011.

[13] Jinhua Du, Jie Jiang, and Andy Way. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 420–429, 2010.

[14] George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, pp. 128–135, 2007.

[15] Atsushi Fujita, Pierre Isabelle, and Roland Kuhn. Enlarging paraphrase collections through generalization and instantiation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 631–642, 2012.

[16] Michel Galley and Christopher D. Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 848–856, 2008.

[17] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pp. 559–578, 2012.

[18] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. Overview of the patent machine translation task at the NTCIR-10 workshop. In *Proceedings of NTCIR-10 Workshop Meeting*, 2013. (in this proceedings).

[19] Zellig Harris. Distributional structure. *Word*, Vol. 10, No. 23, pp. 146–162, 1954.

[20] Xiaodong He. Using word-dependent transition models in HMM-based word alignment for statistical machine translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, pp. 80–87, 2007.

[21] Liang Huang and David Chiang. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 144–151, 2007.

[22] Pierre Isabelle, Cyril Goutte, and Michel Simard. Domain adaptation of MT systems through automatic post-editing. In *Proceedings of the 11th Machine Translation Summit (MT Summit XI)*, pp. 255–261, 2007.

[23] Jie Jiang, Jinhua Du, and Andy Way. Incorporating source-language paraphrases into phrase-based SMT with confusion networks. In *Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Machine Translation*, pp. 31–40, 2011.

[24] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 48–54, 2003.

[25] Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop of Spoken Language Translation (IWSLT)*, 2005.

[26] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.

[27] Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT)*, pp. 120–127, 2007.

[28] Yuval Marton, Chris Callison-Burch, and Philip Resnik. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 381–390, 2009.

[29] Aurélien Max. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 656–666, 2010.

[30] Takashi Onishi, Masao Utiyama, and Eiichiro Sumita. Paraphrase lattice for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL) Short Papers*, pp. 1–5, 2010.

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

[32] Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. Improving translation via targeted paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 127–137, 2010.

[33] Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. Effects of automatic rewriting of source language within a Japanese to English MT system. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pp. 226–239, 1993.

[34] Richard Zens and Hermann Ney. Improvements in phrase-based statistical machine translation. In *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 257–264, 2004.