Classification-Based Approach for Hybridizing Statistical and Rule-Based Machine Translation

Eun-Jin Park, Oh-Woog Kwon, Kangil Kim, and Young-Kil Kim

In this paper, we propose a classification-based approach for hybridizing statistical machine translation and rule-based machine translation. Both the training dataset used in the learning of our proposed classifier and our feature extraction method affect the hybridization quality. To create one such training dataset, a previous approach used auto-evaluation metrics to determine from a set of component machine translation (MT) systems which gave the more accurate translation (by a comparative method). Once this had been determined, the most accurate translation was then labelled in such a way so as to indicate the MT system from which it came. In this previous approach, when the metric evaluation scores were low, there existed a high level of uncertainty as to which of the component MT systems was actually producing the better translation. To relax such uncertainty or error in classification, we propose an alternative approach to such labeling; that is, a cut-off method. In our experiments, using the aforementioned cut-off method in our proposed classifier, we managed to achieve a translation accuracy of 81.5% — a 5.0% improvement over existing methods.

Keywords: Machine translation, hybrid machine translation, automatic labeling, rule-based machine translation, statistical machine translation.

I. Introduction

In the field of machine translation (MT), rule-based approaches have traditionally been used; however, in recent years, statistical approaches have shown promising progress in improving the quality of translations. Rule-based machine translation (RBMT) systems translate source sentences through a deep syntactic analysis, transfer, and generation based on linguistic information [1]–[2]. On the other hand, statistical machine translation (SMT) systems rely on statistical information extracted from bilingual corpuses. Different base knowledge of the two aforementioned systems (SMT and RBMT) causes distinguishable aspects; in addition, hybridizing them can improve the accuracy of translations [3].

One approach to such a hybridization is to train a new classifier to select a translation (*decision*) from among multiple results generated by component MT systems. In such a case, the new classifier's training dataset and corresponding feature extraction method strongly impact upon how the new classifier makes a decision as well as the resulting quality of a decision. A training dataset for such a classifier is often composed of source sentences, reference sentences, and labels (indicating to which of the component MT systems the translation in question belongs).

In previous studies, labels used in training datasets have been produced by utilizing well-known metrics, such as that of the Bilingual Evaluation Understudy (BLEU) [4], which uses a raw bilingual parallel text [5]. Such labels indicate the component MT system from which the translation originated; thus, we may train a classifier, such as our proposed classifier, to assign a label to the best translation result. Ideally, labels output by a classifier are best measured (for accuracy) against human-evaluated labels; however, the limited size of data

Manuscript received Aug. 26, 2014; accepted Jan. 19, 2015.

This work was supported by the ICT R&D program of MSIP/IITP (10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning).

Eun-Jin Park (corresponding author, ejpark@etri.re.kr), Oh-Woog Kwon (ohwoog@etri.re.kr), Kangil Kim (kikim@etri.re.kr), and Young-Kil Kim (kimyk@etri.re.kr) are with the SW & Contents Research Laboratory, ETRI, Daejeon, Rep. of Korea.

available concerning human-evaluated labels has meant researchers have had to resort to auto-evaluation methods, such as BLEU. The BLEU metric tends to favor fluency over accuracy when evaluating a translation. For example, given a source sentence, BLEU counts the occurrence of *n*-grams of the sentence corresponding to good human translation results. Because of the gap between an auto-evaluation method such as BLEU and a human-evaluation method, supervised training results may pose a certain risk. Moreover, BLEU tends to prefer SMT translations over RBMT because BLEU and SMT similarly inherently favor fluency over accuracy; thus, the quality of RBMT translations is easily underestimated.

In this paper, we propose a classification-based approach for hybridizing SMT and RBMT. To manage the aforementioned labeling issue, we devised a cut-off method, whereby given the metric evaluation scores for two translations (one from SMT and one from RBMT), the translation with the highest metric evaluation score is labelled as originating from an SMT system regardless of whether it did in the first place or not — this happens for all translations having a metric evaluation score above a certain threshold (cut-off point). This is because we can be confident that SMT translations will be more accurate than RBMT translations in the cases where they have a high metric evaluation score (due to the similarities between the inherent tendencies of BLEU and SMT).

In addition to our proposed cut-off method, we empirically investigated feature groups to control their effects in learning the classifier for hybridization. We classified the features into six groups according to two criteria. If the internal information of an MT system is to be used (first criterion), then we label groups of features belonging to this class as "glass-box"; if not, then we label groups of features as belonging to a class known as "black-box." The second criterion is related to whether a feature is related to a source sentence, a target sentence, or both.

The remainder of this paper is organized as follows. Section II lists previous related works and a brief explanation about hybrid systems, and Section III shows the limitations of CE-based and classification-based approaches. Next, Section IV describes our classification-based approach and labeling method in detail. Section V then provides the experiment setting and results. Finally, Section VI offers some concluding remarks and a discussion of future works.

II. Related Work

Hybrid systems using a hybridization of MT systems have a greater translation accuracy rate than any system using only one MT system. Such hybrid approaches to improving translation accuracy can be categorized into either classification-based approaches or confidence-estimation (CE) approaches.

1. CE-Based Approaches

CE-based approaches use language models (LMs), alignment information, and linguistic information to estimate the quality of each MT output. They then rank estimate scores and select the highest-ranking score (confidence rank–based approach) as the best translation.

The authors in [6] were the first to attempt to use a CE approach, in which they additionally investigated the use of posterior probabilities of a word graph or *N*-best list to estimate the quality of an MT output. This idea was explored more comprehensively in [7].

In [8], CE estimations were used to re-rank all candidate translations occurring in the *N*-best lists (at the sentence level). The authors then used a combination of CE estimations (at the word level) to reconstruct a single best translation.

In [9], the authors used word-level confidence measures to determine whether a particular translation choice should be accepted or rejected in an interactive translation system. The research in [10] introduced a method for selecting the best translation from a set of translations produced by multiple commercial MT engines using only target trigrams; this method showed a good performance through an evaluation by a human judge on a small dataset. The additional use of an alignment model achieved a 6% improvement over the method that used only a target LM [11]. The proposed hierarchical system combination in [12] used a rudimentary linguistic information–type part of speech, which is helpful for selecting the best translation at the word, phrase, and sentence level.

2. Classification-Based Approaches

A classification-based approach selects the best translation from multiple MT output candidates. In this type of approach, the best-translation selection problem is considered as a classification problem. Most previous researches have adopted supervised machine learning classifiers to resolve such a classification problem [5], [13]–[14]. Ideally, such supervised classifiers should be trained using training datasets that contain human-evaluated labels. As such data is not yet available on a large scale, most previous researches have used training datasets that consist of auto-evaluated labels. Within these training datasets, each source sentence will have a number of associated translations, corresponding to the number of MT engines used. The translations are compared against their human translation (reference) counterparts and an appropriate evaluation score (via a metric) is then assigned to each translation. These evaluation scores can then be ranked, and the

ETRI Journal, Volume 37, Number 3, June 2015
http://dx.doi.org/10.4218/etrij.15.0114.1017

best translation can be identified accordingly. Once the best translation is identified, a label is then assigned to it indicating from which MT the translation originated.

In [13], for automatically constructing a training set, the word-level Levenshtein distance was used as an evaluation metric, and three classification classifiers, each using a different classification algorithm (Naïve Bayes, linear regression, and support vector machine (SVM)), were used for feature selection.

In [5], when constructing their training dataset, the author applied Meteor, National Institute of Standards and Technology (NIST), and BLEU to each MT output to estimate a ranking at the sentence level. The author could then rank the resulting estimations to find the best translation. For pairwise system comparisons, binary SVM classifiers were trained from the decomposed training dataset, and the MT engines within the author's proposed hybrid MT system participated in a round-robin playoff to find the single best output.

3. Combining CE-Based and Classification-Based Approaches

In [14], a sentence-level BLEU was introduced, and three approaches were considered — confidence rank–based approach, classification-based approach, and a combination of the two. The experiment in [14] showed that the confidence rank–based approach outperformed the classification-based approach under SVM. In addition, the combination of the two approaches showed a similar performance to the confidence rank–based approach.

III. Limits of Hybrid Approaches

The two most widely used approaches, confidence rank–based approach and classification-based approach, have limitations when applied to MT hybridizations. Confidence estimation methods are designed to evaluate only a single translation in isolation; thus, they do not consider other MT translation results when assigning an evaluation score [10]–[12]. Moreover, they are biased toward information or features pertaining to a particular MT system. A classification-based approach to hybrid MT has a limitation in terms of translation accuracy. A classifier for a hybrid MT system is expected to select, from a human perspective, the *best* translation from multiple candidate translations; hence, the classifier's training dataset should contain labels that have been evaluated by humans and not by other methods. In reality, due to insufficient quantities of datasets containing human-evaluated labels, the labeling process is imitated by using auto-evaluation metrics, such as BLEU and NIST [5], [13]–[14]. However, use of such auto-evaluation metrics is obviously not going to be as accurate as a human evaluation. BLEU is a good example of this. It evaluates an MT translation by comparing it against a reference translation from the training set and counting the number of *n*-grams (from reference sentence). If the number of reference sentences per source sentence is small, then the metric may not be able to evaluate all words in a given translation; this is because there may be some words in the given translation that don't have any corresponding words in the reference translations. Thus, such words would be ignored in the metric when evaluating the translation quality [14].

SMT systems aim at learning non-linguistic translation knowledge rather than statistical translation knowledge from massive amounts of human translations, and by working on this knowledge, they then aim to imitate a human translation. On the contrary, RBMT systems aim at implementing translation process rules based on linguistic theories, as well as working on the sophisticated linguistic rules constructed by experts. Because of the different paradigms between SMT and RBMT, their strengths and shortcomings may be different from the perspective of human translation. The method of auto-evaluation is based on statistical information, which favors SMT systems and underestimates RBMT systems.

A combination of CE-based and classification-based approaches may be complementary in reducing the limitations of SMT and RBMT systems. A confidence rank–based approach is limited by the number of possible features it can employ to estimate the quality of a translation. Classification-based approaches, by comparison, can use more features and are thus less restricted as a result of doing so. We think that by combining the two approaches, we can obtain a better performance than when using them individually.

In our method, when labeling the classifier's training set, we labeled the high BLEU scores with an SMT label and the remaining scores with an RBMT label. Because of the mechanism of RBMT, it is thought that RBMT provides a more accurate translation in the cases of such low BLEU scores.

In this paper, we propose a framework merging the confidence-based and classification-based methods in Section IV. For a clearer analysis, we focus on combining SMT and RBMT systems, which have been widely used in MT research.

IV. CE-Based Training of a Classifier for Combining MTs

1. Classification of Confidence Rank Group for Combination

To relax limitations, we propose the use of a classifier for the combination and train it from data guided by the confidence rank. As shown in Fig. 1, it is a general frame of combining systems based on a classifier. The distinguishing point is

controlling the training data. Given a set of sentences for training, we evaluate their confidence ranks and sort them. We then truncate them to determine their labels as outputs of the classifier. This classifier predicts the output given the feature vectors extracted from different information sources denoted by G_i.

The truncation strategy simplifies the problem of evaluating the accuracy of a translation output. Thus, this allows hybrid systems to predict binary-valued classes, rather than accurate continuous value estimation, which reduces the complexity of the algorithm. Compared with estimating an accurate confidence rank, it predicts only their classes. Learning the main inclination rather than a sophisticated difference, the simplified model is expected to reduce errors caused by the gap between ranking and a human evaluation.

A truncation method for combing the SMT and RBMT is written as the following equation:

$$f(\mathbf{f};s)=\begin{cases} l_{\mathrm{SMT}} & \text{if } c(s_{\mathrm{SMT}})\geq\theta_c, \\ l_{\mathrm{RBMT}} & \text{otherwise,} \end{cases} \quad (1)$$

where θ_c is the threshold for a confidence measure c to determine a label l_{SMT} or l_{RBMT}, and $c(s_{\mathrm{SMT}})$ is the confidence value of sentence s. Given the sentence, this function predicts a label for the input feature vector $\mathbf{f}$ related to s and other resources.

A classifier trained by this data frame behaves differently from a direct decision made by a confidence rank. In selecting high-rank sentences, they will be similar for the observed data of the training set. However, the prediction of the unobserved data differs. In our framework, a decision on the unobserved data is flexibly affected by the set of features. However, a direct decision model based on the confidence rank leads to a different decision distribution over unobserved data because it uses a fixed model for all data. This flexibility leads to robustness of the combining system to the wrong feature selection. In particular, the limitation of using isolated information in evaluating the confidence rank is easily solved by adding features related to other MTs. To show this strength of our frame, we investigated the available feature sets.

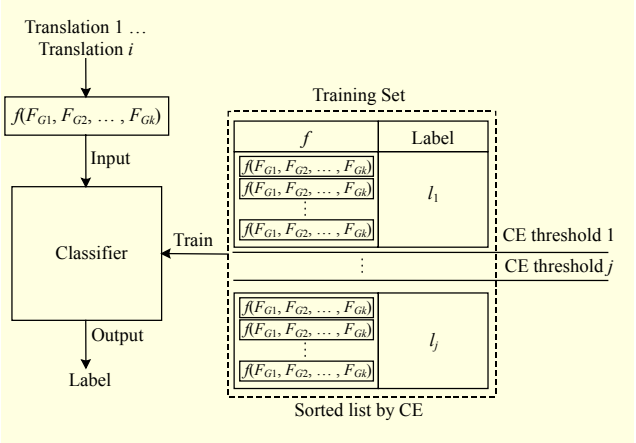

Fig. 1. System combination.

2. Features for MT Combination

The features used in the MT system combination can be divided into black-box and glass-box features [12]. Black-box features, such as perplexity numbers of the *n*-gram LM and the length of the input/output sentences, are derived from the input/output string. Therefore, the black-box features can be applied to a large variety of MT approaches from the SMT to the RBMT. On the other hand, glass-box features are extracted based on the internal detailed information of each MT system. The typical examples of glass-box features are derived from part-of-speech (POS) of words; the syntactic tree; word and phrase translation pairs; and so on. Some glass-box features cannot be extracted from an MT system that does not provide its related internal information; thus, glass-box features are MT-system dependent. Unlike the division between a black-box and glass-box, the features can also be separated into source side–based features, target side–based features, and dual-sided-based features according to their origin.

The features we used in this paper are divided into six groups: source side–based black-box features (SBFs), target side–based black-box features (TBFs), dual-sided-based black-box features (DBFs), source side–based glass-box features (SGFs), target side–based glass-box features (TGFs), and dual-sided-based glass-box features (DGFs).

A. SBFs

SBFs are directly extracted from the source sentences and source LMs. If the source LMs are trained from the source sentences of an SMT training set, then the information of the LM is useful to determine whether the SMT can translate an input sentence well or poorly. In this paper, we proposed six SBFs as follows:

- The length of the source sentence.
- Five perplexities of the source sentence computed by 1-to-5-gram source LMs.

B. TBFs

As the SBFs, we propose TBFs extracted from the target sentences and target LMs. This is defined as follows:

- The length of each candidate output.
- Five perplexities of each candidate output computed by 1-to-5-gram target LMs.

In combining the SMT and RBMT, the number of candidate

ETRI Journal, Volume 37, Number 3, June 2015
http://dx.doi.org/10.4218/etrij.15.0114.1017

outputs is two, and a total of 12 features are extracted.

C. DBFs

We derive six DBFs from a comparison of the source sentence and each MT output as follows:

- The length difference between the given source sentence and each candidate translation sentence.
- Five differences between five perplexities of the source sentence computed using 1-to-5-gram source LMs and five perplexities of the target sentence computed by 1-to-5-gram target LMs.

Thus far, we have defined 30 black-box features consisting of six SBFs, 12 TBFs, and 12 DBFs.

D. SGFs

We can often see that a certain MT system translates sentences well or poorly according to their written style, sentence structures, and vocabularies. To roughly predict the characteristics of input sentences, we extracted five SGFs, which are derived from POSs of the words. These five features are the numbers of morphemes, verbs, nouns, postpositions, and endings. The number of endings was used for an agglutinative language, Korean, which is treated as the source language in this paper.

We also extracted two SGFs from the translation units (words and phrases), which each MT system uses to search for the target equivalents during the translation process. One is the number of translation units, and the other is the average perplexities of the translation units computed by the source LMs.

E. TGFs

We used three TGFs derived from the counterparts of the translation units — the count of unknown equivalents, the count of known equivalents, and the average perplexities of the equivalents (target words and phrases) computed by the target LMs.

F. DGFs

Two DGFs are derived from the alignments between source sentences and their translation counterparts. Both DGFs are the averages of the probabilities of the translation units that are translated to their final target equivalents. One DGF was based on the probabilities provided by each MT system; thus, the probabilities might be estimated differently depending on the probability estimation function and the samples used in the system. The other DGFs were computed based on new probabilities estimated from the same parallel dataset to eliminate the difference. To combine the SMT and RBMT, we extracted 17 glass-box features comprised of seven SGFs, six TGFs, and four DGFs. The total number of features in this paper is 47.

V. Experiments

We empirically evaluate the effects of the proposed combiner frame on a practical translation performance.

1. System Setting

A. Hybrid Architecture

The whole system for our translation experiment is composed of three main components. Independently built SMT and RBMT provide two translation results from a given input sentence. A combiner based on an SVM then extracts various features from all available resources explained in Section IV, including the input sentence, dictionary, LM, and computed statistical information. From the input feature values, the SVM determines which MT generates the best translation.

B. RBMT

We used an RBMT system based on a structure transfer approach for Korean to Chinese translation. It consists of a linear-chain conditional random fields (CRF)-based Korean POS tagger [15], a graph-based dependency parser, a syntactic tree-to-tree transfer based on about 150,000 rules built by human experts, and a Chinese generator. The RBMT system was developed by us with the aim of focusing on the multilingual expansibility of RBMT using knowledge learning. The details of our RBMT are out of the scope of this paper.

C. SMT

We adopted the MOSES package to obtain the SMT results. We use a phrased-based SMT of the package and set the default value for the parameters [16]. We trained it from a training set of the ETRI travel dialog KC dataset consisting of 2,124,196 Korean–Chinese parallel sentences regarding travel dialogues, collected by the Natural Language Processing Department of the Electronics and Telecommunications Research Institute. To improve the performance of Korean to Chinese translation, we use a linear-chain CRF model [16] for Korean word segmentation and the Stanford Chinese Segmenter [17] for Chinese word segmentation.

D. Combiner Model

We use libSVM [18], an SVM toolkit, for building the

combiner. To build a binary SVM classifier, we set the SVM type to C-support vector classification and the kernel function type to a radial basis function. The values of the parameters are as follows: 32 for the cost (denoted by c in the package), 0.5 for gamma (g), and 0 for shrinking (h). This setting is tuned by a script (grid.py) supported by libSVM [19]. To train them, we use the following two training sets:

- SMT-dependent training dataset: 20,000 parallel sentences randomly selected from our SMT training set.
- SMT-independent training dataset: 20,000 parallel sentences randomly selected from the test set of the ETRI travel dialogue KC dataset consisting of 51,510 Korean–Chinese parallel sentences irrelevant to the SMT training set.

Table 1 shows further details of the SMT training dataset and two training datasets for the SVM combiner.

To construct the training sets for selecting the best translation results between SMT and RBMT, we translated Korean sentences from these training sets into Chinese sentences using the above-mentioned SMT and RBMT systems. The training datasets consist of Korean sentences, their Chinese sentences translated by humans, their translation results by the RBMT, and their translation results by the SMT.

Table 1. Detailed information of training sets.

Training set	Components	# of sentences	Total # of words	Avg. # of words per sentence
SMT-training dataset	Source sentences	2,124,196	17,212 K	8.1
	Human translations		15,011 K	7.1
SMT-dependent training dataset for SVM	Source sentences	20,000	174 K	8.7
	Human translations		149 K	7.5
	SMT outputs		146 K	7.3
	RBMT outputs		135 K	6.8
SMT-independent training dataset for SVM	Source sentences	20,000	228 K	10.3
	Human translations		177 K	8.0
	SMT outputs		164 K	7.4
	RBMT outputs		180 K	8.1

Table 2. Confidence values of translated results from training sets.

		BLEU	NIST	METEOR
SMT-dependent training set	SMT	0.6788	12.0317	0.4934
	RBMT	0.0168	1.1502	0.0707
SMT-independent training set	SMT	0.1224	4.9024	0.1994
	RBMT	0.0730	4.2167	0.2319

We used two training sets for a fairer comparison. Table 2 shows the confidence values of the translation results of the training sets. As the results show, a subset of the SMT increases the measures of the SMT outputs, and thus it may cause a bias in selecting one of them through a combiner.

To reduce noise and improve scaling, for the SVM, we excluded sentences generating outliers of the feature inputs. The criteria for detecting them are defined as follows:

$$|\mu_i - x_i| > w\sigma_i, \tag{2}$$

where μ_i and σ_i are the mean and standard deviation of the ith feature, respectively, and w is the criterion weight, which is set to a value of three. If a feature of a given sentence satisfies this condition, then we omit it.

We set the available feature groups in Section IV. Among the feature groups, the LM-related features and perplexity of the source and target LM are extracted from the source and target LM using the publically available KenLM toolkit [20].

2. Evaluation Setting

We used test sets for the evaluation of the translation quality collected from two different sources. One set is made up of 100 sentences from the travel domain, and the other is made up of 225 sentences from the GenieTalk[1)] application log, which is denoted by gtalk in the following sections. The travel set has more formal and cleaner expressions than the gtalk set. This set is composed of source sentences, their translated results from two MTs, and their translation quality as evaluated by humans. For the human evaluation, we adopt the criteria shown in Table 3 for scoring the translation accuracy, which were used in preliminary works [21]. In our human evaluation, five professional translators evaluated the results of the SMT and RBMT. Ruling out the highest and lowest scores, three scores

Table 3. Scoring criteria for translation accuracy.

Score	Criterion
4	Meaning of a sentence is perfectly conveyed.
3.5	Meaning of a sentence is almost perfectly conveyed except for some minor errors (e.g. wrong article and stylistic errors).
3	Meaning of a sentence is almost conveyed (e.g. some errors in target word selection).
2.5	Simple sentence in a complex sentence is correctly translated.
2	Sentence is translated phrase-wise.
1	Only some words are translated.
0	No translation.

1) https://play.google.com/store/apps/details?id=kr.re.etri.saytran.phone

ETRI Journal, Volume 37, Number 3, June 2015
http://dx.doi.org/10.4218/etrij.15.0114.1017

Table 4. Evaluation results of evaluation sets.

Translation quality metrics	Travel evaluation set		Gtalk evaluation set	
	SMT	RBMT	SMT	RBMT
Translation accuracy	76.3%	77.1%	71.9%	70.6%
BLEU	0.1426	0.1407	0.1091	0.0965
NIST	3.7451	3.9279	3.4208	3.6241
METEOR	0.2752	0.2919	0.2589	0.2755

were used for the translation accuracy evaluation. The human-evaluated translation accuracy is defined as follows:

$$\text{Translation accuracy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{5}\sum_{j=1}^{5}\left(\frac{1}{4}\text{score}_j\right)\right)\times 100, \quad (3)$$

where n is the number of test sentences, and score_j is the score evaluated by professional translator j.

Table 4 shows the translation accuracy of the human evaluation, BLEU, NIST, and METEOR for the results of the SMT and RBMT systems experimented on using two evaluation sets. In the task for combining the results of the SMT and RBMT, the upper bound of the translation accuracy is 89.6% for the evaluation of the travel set and 83.4% for the gtalk set. The gtalk set is three words shorter than the travel evaluation set, and thus it is disadvantageous for BLEU matching using only a 4-gram.

Beyond evaluating the accurate score of the human evaluation, we measure the classification accuracy of the predicting labels for evaluating the performance of our classifier. For this, we use the following measure:

$$\text{Classification accuracy} = \frac{|\text{true positive}| + |\text{true negative}| + |\text{DC}|}{|\text{prediction trial}|}, \quad (4)$$

where DC is the number of sentences equally scored by humans. This is a normal accuracy definition with the exception of the DC term; that is, indicating pairs of translated results obtaining the same score in a human evaluation.

3. Methods for Comparison (Ranking vs. Truncation)

To evaluate the comparative performance of our method, we select a ranking method to generate label data for training the SVM combiner, which is defined as follows:

$$f(\mathbf{f}; s_{\text{RBMT}}, s_{\text{SMT}}, s) = \begin{cases} \text{SMT} & \text{if } c(s_{\text{RBMT}}) > c(s_{\text{SMT}}), \\ \text{RBMT} & \text{otherwise,} \end{cases} \quad (5)$$

where s_i is a translated result for sentence s translated by MT i (i = SMT or RBMT). This equation selects which MT generates a better translation by comparing the confidence scores for the translated results.

We compare this ranking method with our truncation approach defined in (1) for three confidence measures: BLEU, NIST, and METEOR. To investigate the change based on threshold θ_c, we evaluate the performance by increasing the training sentences by 5% from 5% to 95% and changing the confidence value θ_c corresponding to the division.

A. Performance

We investigated the performance by measuring the translation accuracy evaluated by humans (T-accuracy), classification accuracy (C-accuracy), and BLEU score for two labeling methods: ranking and truncation. The labeling results are shown in Table 5 for the SMT-independent training set and Table 6 for the SMT-dependent training set. Confidence metrics are used for the calculation of $c(s)$ for ranking and truncation.

In Table 5, the best human evaluation score is 81.5% for the BLEU truncation method, which is larger than a single SMT by 5.2% and single RBMT by 4.4%. The accuracy of the best truncation is improved by 3.0% compared to the single SMT and 5.0% compared to the single RBMT. The performance improvement of the truncation method is observed in all comparison cases: combination of three measures, different evaluation sets, and evaluation scores. Table 6 shows the hybrid performance of the SMT-dependent training set. The overall performance is still improved in the best truncation, although it shows a lower accuracy than ranking using BLEU for the

Table 5. Hybrid performance of training combiner through SMT-independent set.

Confidence metrics	Combination method	Travel evaluation set			Gtalk evaluation set		
		T-accuracy (%)	C-accuracy (%)	BLEU	T-accuracy (%)	C-accuracy (%)	BLEU
NIST	Ranking	77.0	64.0	0.1300	70.8	68.9	0.1024
	Best truncation	80.5	67.0	0.1635	75.3	69.3	0.1179
BLEU	Ranking	78.6	65.0	0.1574	72.0	69.3	0.1093
	Best truncation	81.5	69.0	0.1535	75.2	79.1	0.1189
METEOR	Ranking	73.4	54.0	0.1335	70.9	70.7	0.0999
	Best truncation	77.9	64.0	0.1577	72.1	70.2	0.1140
—	Single SMT	76.3	61.0	0.1426	71.9	71.9	0.1091
	Single RBMT	77.1	59.0	0.1407	70.6	70.6	0.0965

Table 6. Hybrid performance of training combiner through SMT-dependent set.

Confidence metrics	Combination method	Travel evaluation set			Gtalk evaluation set		
		T-accuracy (%)	C-accuracy (%)	BLEU	T-accuracy (%)	C-accuracy (%)	BLEU
NIST	Ranking	76.3	61.0	0.1426	71.9	68.9	0.1091
	Best truncation	80.3	68.0	0.1491	72.7	70.2	0.1260
BLEU	Ranking	76.3	61.0	0.1426	71.9	68.9	0.1091
	Best truncation	80.8	71.0	0.1465	72.5	72.0	0.1239
METEOR	Ranking	77.1	59.0	0.1407	70.6	64.9	0.0965
	Best truncation	80.0	67.0	0.1510	72.4	69.8	0.1213
—	Single SMT	76.3	61.0	0.1426	71.9	71.9	0.1091
	Single RBMT	77.1	59.0	0.1407	70.6	70.6	0.0965

truncation of the travel set. The performance improvement of all sets implies that a truncation can generate better combining results, especially in a human-translation evaluation.

B. Feature Group Analysis

We evaluated the performance of each feature group defined in Section IV. The results are shown in Table 7 for the travel evaluation set and Table 8 for the gtalk evaluation set. Using all features, the hybrid system using a truncation shows a 0.5% improvement in translation accuracy (81.5%) compared to the best translation accuracy (81.0% using only TGFs) of other feature combinations. In contrast, the ranking shows a 3.3% decrease when it uses all features. Even in the evaluation of gtalk set, the same result is observed, as shown in Table 8. The truncation labeling method shows the best performance when using all feature groups for both evaluation sets. Compared to truncation, the ranking method shows that a single feature group may show a better performance but is inconsistent in the evaluation sets.

VI. Conclusion

In this paper, we proposed a classification-based hybridization to select the best between translation results of a statistical machine translation (SMT) and a rule-based machine translation (RBMT). This approach has a limit in hybridizing an SMT with an RBMT, because such measures are designed for evaluating the fluency, and an SMT has an advantage over an RBMT. This may cause a biased preference toward an SMT.

Table 7. Feature group performances evaluated for travel set.

Feature Group	Truncation BLEU		Ranking BLEU	
	T-accuracy (%)	C-accuracy (%)	T-accuracy (%)	C-accuracy (%)
All	81.5	69.0	78.6	65.0
SBFs	77.0	60.0	77.5	63.0
BBFs	77.6	61.0	76.3	55.0
SGFs	77.1	60.0	77.9	63.0
TGFs	81.0	71.0	76.3	61.0
BGFs	76.6	62.0	76.3	61.0
TBFs	76.0	60.0	80.8	69.0
Black-box	75.4	56.0	81.3	69.0
Glass-box	78.5	64.0	78.9	67.0
Source-side	78.5	63.0	81.3	69.0
Target-side	76.3	60.0	80.3	67.0
Both-side	76.8	63.0	80.0	70.0

Table 8. Feature group performances evaluated for gtalk set.

Feature Group	Truncation BLEU		Ranking BLEU	
	T-accuracy (%)	C-accuracy (%)	T-accuracy (%)	C-accuracy (%)
All	80.8	71.0	76.3	61.0
SBFs	71.5	68.4	71.9	69.3
BBFs	71.6	67.6	69.5	60.9
SGFs	72.2	71.1	71.9	68.9
TGFs	74.1	76.0	71.9	68.9
BGFs	71.9	68.9	73.2	71.6
TBFs	71.5	68.4	69.5	62.2
Black-box	72.6	70.2	69.7	63.1
Glass-box	72.6	71.1	73.0	72.4
Source-side	72.2	70.7	71.9	70.2
Target-side	72.1	72.4	71.7	71.6
Both-side	73.0	71.1	72.0	66.7

To manage this issue, we proposed a method to cut off an uncertain translation from a label prediction. In our experiment, this generation method improved the translation accuracy by 5.0%, compared to labeling through competition.

The features used to determine a better translation should be deeply related to both the source sentence and other participating MT outputs; however, the dependency between features has not been deeply analyzed thus far in the MT literature; and assuming their independence may be risky. We investigated a better combination of feature sets divided into

ETRI Journal, Volume 37, Number 3, June 2015
http://dx.doi.org/10.4218/etrij.15.0114.1017

six groups: SBFs, TBFs, DBFs, SGFs, TGFs, and DGFs. In our experiments, using all groups showed the best translation quality and classifier accuracy, as compared with other feature combinations. Overall, the proposed classification-based hybrid approach achieved an 81.5% translation quality, while an RBMT and phrase-based SMT showed a translation quality of 77.1% and 76.3%, respectively. The improvement in the truncation in an SMT-dependent set is expected to have a practical benefit in constructing a hybrid system because it seems to guarantee a better performance in reusing the SMT training set for the combiner training. This implies that we can reduce the additional cost for building a new set for the combiner, resolving the main issue in MT research. We leave an analysis of the generalization as a future work.

References

[1] Y.A. Seo, S.K. Park, and K.S. Choi, "Structural Disambiguation of Korean Adverbs Based on Correlative Relation and Morphological Context," *ETRI J.*, vol. 28, no. 6, Dec. 2006, pp. 803–806.

[2] S.I. Yang et al., "Noun Sense Identification of Korean Nominal Compounds Based on Sentential Form Recovery," *ETRI J.*, vol. 32, no. 5, Oct. 2010, pp. 740–749.

[3] A. Eisele et al., "Hybrid Machine Translation Architectures within and beyond the EuroMatrix Project," *Proc. Annual Conf. European Association Mach. Transl.*, Hamburg, Germany, Sept. 22–23, 2008, pp. 27–34.

[4] K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proc. Association Comput. Linguistics*, Philadelphia, PA, USA, July 2002, pp. 311–318.

[5] C. Federmann, "Hybrid Machine Translation Using Joint, Binarised Feature Vectors," *Proc. Conf. Association Mach. Transl. Americas*, San Diego, CA, USA, Oct. 2012, pp. 113–118.

[6] N. Ueffing, K. Macherey, and H. Ney, "Confidence Measures for Statistical Machine Translation," *Mach. Transl. Summit*, New Orleans, LA, USA, Sept. 2003, pp. 394–401.

[7] J. Blatz et al., "Confidence Estimation for Machine Translation," *Int. Conf. Comput. Linguistics*, Geneva, Switzerland, Aug. 23–27, 2004, pp. 315–321.

[8] C.B. Quirk, "Training a Sentence-Level Machine Translation Confidence Measure," *Int. Conf. Language Resources Evaluation*, Lisbon, Portugal, May 26–28, 2004, pp. 825–828.

[9] N. Ueffing and H. Ney, "Application of Word-Level Confidence Measures in Interactive Statistical Machine Translation," *Annual Conf. European Association Mach. Transl.*, Budapest, Hungary, May 30–31, 2005, pp. 262–270.

[10] C. Callison-burch and R.S. Flournoy, "A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines," *Proc. Mach. Transl. Summit VIII*, Sept. 18–22, 2001, pp. 63–66.

[11] Y. Akiba, T. Watanabe, and E. Sumita, "Using Language and Translation Models to Select the Best among Outputs from Multiple MT systems," *Proc. COLING*, Aug. 26–30, 2002, pp. 8–14.

[12] F. Huang and K. Papineni, "Hierarchical System Combination for Machine Translation," *Proc. Empirical Methods Natural Language Process*, Prague, Czech Republic, June 2007, pp. 277–286.

[13] E. Avramidis, "DFKI System Combination with Sentence Ranking at ML4HMT-2011," *Proc. Int. Workshop Using Linguistic Inf. Hybrid Mach. Transl. Shared Task Applying Mach. Learning Techn. Optimise Division Labor Hybrid Mach. Transl.*, Barcelona, Spain, Nov. 18, 2011.

[14] R. Soricut and S. Narsale, "Combining Quality Prediction and System Selection for Improved Automatic Translation Output," *Proc. Workshop Statistical Mach. Transl. Association Comput. Linguistics*, Jeju, Rep. of Korea, July 2012, pp. 163–170.

[15] S.H. Na, C.H. Kim, and Y.K. Kim, "Two-Stage Compound Morpheme Segmentation in CRF-Based Korean Morphological Analysis," *Annual Conf. Human Cognitive Language Technol.*, Seoul, Rep. of Korea, Oct. 2007, 2013, pp. 13–17.

[16] P. Koehn et al., "Moses: Open Source Toolkit for Statistical Machine Translation," *Proc. Annual Meeting ACL Interactive Poster Demonstration Sessions Association Comput. Linguistics*, Prague, Czech Republic, June 23–30, pp. 177–180.

[17] H. Tseng et al., "A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005," *Proc. SIGHAN Workshop Chinese Language Process.*, Jeju, Rep. of Korea, Oct. 14–15, 2005, pp. 168–171.

[18] C.-C. Chang and C.-J. Lin, "libSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, pp. 27:1–27:27.

[19] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Apr. 15, 2010.

[20] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," *Proc. Workshop Statistical Mach. Transl. Association Comput. Linguistics*, Edinburgh, UK, July 30–31, 2011, pp. 187–197.

[21] O.-W. Kwon et al., "Customizing an English-Korean Machine Translation System for Patent/Technical Documents Translation," *Proc. Pacific Asia Conf. Language, Inf. Comput.*, Hong Kong, China, Dec. 3–5, 2009, pp. 718–725.

Eun-Jin Park received his BS and MS degrees in computer science from the Korea Maritime and Ocean University, Busan, Rep. of Korea, in 2006. He is currently pursuing his PhD degree in computer software at the Korea University of Science and Technology, Daejeon, Rep. of Korea. Since 2010, he has been working for the Electronics and Telecommunications Research Institute. His current research interests include natural-language processing, dialogue understanding, and machine translation.

Oh-Woog Kwon received his MS degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 1995 and his PhD degree in computer engineering from Pohang University of Science and Technology, Rep. of Korea, in 2001. He joined the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, in 2004. His major research interests include natural-language processing, dialogue processing, and machine translation.

Kangil Kim received his BS degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2006 and his PhD degree in electrical engineering and computer science from Seoul National University, Rep. of Korea, in 2012. Since 2013, he has been working as a senior researcher with the Natural-Language Processing Section, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests include artificial intelligence, evolutionary computation, machine learning, and natural-language processing.

Young-Kil Kim received his MS and PhD degrees in electronics and telecommunications from Hanyang University, Seoul, Rep. of Korea, in 1993 and 1997, respectively. He has been a principal member of the engineering staff and section leader of the Natural-Language Processing Section, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests include natural-language processing, dialogue understanding, and machine translation.