Improved Spoken Language Translation Using \( N \)-best Speech Recognition Hypotheses

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Frank Soong, Wai-Kit Lo, Taro Watanabe, Eiichiro Sumita

ATR Spoken Language Translation Laboratories
2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto, Japan 619-0288

Abstract

We intended to demonstrate the effect of using \( N \)-best speech recognition hypotheses for improving speech translation performance. A log-linear model, which integrated features from speech recognition and statistical machine translation, was used to rescore the translation candidates. Model parameters were estimated by optimizing an objectively measurable but subjectively relevant translation quality metric. Experimental results have shown that the proposed \( N \)-best approach improved translation quality over the conventional single-best approach. The improvements were confirmed consistently by several automatic translation evaluation metrics.

1. Introduction

Spoken language translation, due to the non-robustness of speech recognition, usually cannot achieve the same level of translation performance as that achieved by perfect text input. The inevitable recognition errors are one of the causes for translation degradation in the current spoken language translation. Hence, a high performance speech translation system should be designed to alleviate impairment to translations due to speech recognition errors.

One of the major research concerns with regards to speech translation is to implement a tight integration of speech recognition and machine translation to achieve overall system optimization. Several architectures for speech translation have been proposed so far. In [1] a coupling structure was proposed where acoustic model and translation model are in effect harmoniously. In [2] a unified structure was proposed where the maximum entropy approach is applied to build the speech translation model directly. But the maximum entropy algorithm is very hard to be implemented in complicated task, such as speech translation.

In this paper we use the log-linear model to integrate features derived from speech recognition and machine translation to optimize speech translation performance. Easy implementation and effective integration are the merits of the model. The log-linear model has been studied in speech recognition and machine translation research separately [3] [4]. But no result has been reported in speech translation.

As a preliminary investigation we start to test our approach by using \( N \)-best speech recognition hypotheses. It is widely accepted in the speech recognition community that making use of \( N \)-best recognition hypotheses can improve speech recognition accuracy, for example, by means of building advanced language models [5]. Because \( N \)-best hypotheses carry much more information than the single-best, they can improve speech translation performance if used properly.

A statistical machine translation (SMT) system were used in our experiments. SMT has been shown to be a more effective approach in dealing with spoken language translation than text-based translation [6]. Features derived from speech recognition conveyed in \( N \)-best hypotheses and SMT are integrated by the log-linear model. We found that this model provided a seamless way to combine those features.

Using \( N \)-best recognition hypotheses in speech translation can be found in [7], where \( N \)-best recognition hypotheses were post-processed into a single-best hypothesis by grammar-based analysis before translation. But in fact it is the single-best translation and does not achieve joint optimization of speech recognition and machine translation. In [1] \( N \)-best translation was mentioned briefly but no experimental results were reported.

In section 2 we introduce our speech translation system and the feature-based log-linear models. In section 3, we describe the optimization algorithm used to find the parameters of the log-linear model. In section 4 we demonstrate the effectiveness of the new approach in speech translation experiments. In the last section we present our conclusion and propose some future work.

2. Feature-based Log-linear Models in Speech Translation

The statistical speech translation system in this study is illustrated in Fig. 1. It consists of two major cascaded components: an automatic speech recognition (ASR) module and a statistical machine translation (SMT) module. Additionally, a third module, ‘Rescore’, is added to the system as a key component of our approach. Features derived from ASR and SMT are used together to rescore translation candidates.

Without loss of generality, in this paper we used Japanese-to-English translation as an example to explain the generic speech translation process. Let \( X \) denote acoustic observations of a Japanese utterance, typically a sequence of short-time spectral vectors received at a frame rate of every centi-second. It is first recognized as a Japanese sentence, \( J \). The recognized sentence is then translated into a corresponding English sentence, \( E \).

The conversion from \( X \) to \( J \) is performed in the ASR module. Based on Bayes’ rule, the conditional probability \( P(J|X) \) can be written as

\[
P(J|X) = \frac{P_{am}(X|J)P_{lm}(J)}{P(X)}
\]

where \( P_{am}(X|J) \) is the acoustic model likelihood of the observations given the recognized sentence \( J \); \( P_{lm}(J) \), the Japanese language model probability; and \( P(X) \), the probability of all acoustic observations.

\[
P_{am}(X|J) = \prod_{t} P_{am}(x_{t}|J)
\]

where \( x_{t} \) is the 13-dimensional feature vector of the acoustic observations.
Figure 1: Current framework of speech translation

By the ASR a set of $N$-best hypotheses are generated, $J_i^N = \{J_1, J_2, \cdots, J_N\}$, and each $J_i$ is determined by

$$J_i = \arg \max_{J \in \Omega_i} P_{asr}(X|J) P_{lm}(J)$$

where $\Omega_i$ is the set of all possible sentences in the Japanese language excluding all higher ranked $J_k$’s, $1 \leq k \leq i - 1$.

The statistical machine translation module, SMT, performs the conversion from $J$ to $E$. According to the IBM statistical machine translation model, it functions to search for the best sentence $E$ such that

$$\hat{E} = \arg \max_{E \in \Omega} P(E|J) = \arg \max_{E \in \Omega} P(J|E) P(E)$$

where $P(J|E)$ is a translation model characterizing the correspondence between $E$ and $J$; $P(E)$, the English language model probability.

In IBM Model 4, the translation model $P(J|E)$ is further decomposed into four sub-models:

- Lexicon model - $t(j|e)$
- Fertility model - $n(\phi|e)$
- Distortion model - $d$
- NULL translation model - $p_i$

In the descriptions above we listed seven features: two from ASR ($P_{asr}(X|J), P_{lm}(J)$) and live from SMT ($P(E), t(j|e), n(\phi|e), d, p_i$).

The third module in Fig. 1 is of rescoring translation hypotheses by using a feature-based log-linear model. All speech translation candidates output by statistical machine translation modules are re-evaluated by this model integrating all relevant features from speech recognition and machine translation modules, and the best translation candidate with the highest score is found.

The log-linear model used in our experiments, $P(E|X)$, is

$$P_{\lambda}(E|X) = \frac{\exp \left( \sum_{E'}^M \sum_{i=1}^N \lambda_i f_i(X, E') \right)}{\exp \left( \sum_{E'}^M \sum_{i=1}^N \lambda_i f_i(X, E') \right)} \quad \Lambda = \{\lambda_i^M\}$$

In Eq. 1, $f_i(X, E)$ is the logarithm value of the $i$-th feature, $\lambda_i$ is the weight of the $i$-th feature.

In addition to the above seven features, the following features were also incorporated:

- Part-of-speech language models
- Length model $P(|l|E, J)$
- Jump weight
- Phrase matching score
- Dynamic phrase matching score

Detail explanation about the above features can be found in [8]. In all, we used $M(=12)$ different features. In next section we describe our approach to optimize objective translation metrics for the weights of the features, $\lambda_i^M$.

4. Experiments

4.1. Corpus & Model Training

The data used in the experiments was extracted from the Basic Travel Expression Corpus (BTEC) [12], consisting of sentences from travel guidebooks and tour conversations. The corpus was designed to assist the development of multiple language speech-to-speech translation systems. It contains four

3. Optimize Translation Metrics

We approximate (Eq. 1) by ignoring the denominator because the normalization is applied equally to every hypothesis. Hence, the best translation $\hat{E}$ over all possible translations, $E$, is

$$\hat{E} = \arg \max_{E \in \Omega} \sum_{i=1}^M \lambda_i log P_i(X, E)$$

where we write features, $f_i(X, E)$, explicitly in their logarithms, $log P_i(X, E)$.

The efficiency of the model in Eq. 2 depends upon the parameter optimization of the set $\lambda_i^M$ with respect to some objectively measurable but subjectively relevant metrics.

Assume that we have $L$ speech utterances and for each utterance, we generate the top $N$ speech recognition hypotheses. And for each recognition hypothesis, we generate $K$ English translation hypotheses. Therefore, the $l$-th input speech utterance makes $N \times K$ translations, $C_l = \{E_{l1}, \cdots, E_{lN \times K}\}$. All $L$ speech utterances generate $L \times N \times K$ translations in total.

Our goal is to minimize the translation “distortion” between the translated sentences, $\hat{E}$ and the reference translations, $R$.

$$\lambda_i^M = \text{optimize } D(\hat{E}, R)$$

where $\hat{E} = \{\hat{E}_1, \cdots, \hat{E}_L\}$ is a set of translations of all utterances. The translation $E_l$ of the $l$-th utterance is produced by Eq. 2, where $E \in C_l$.

Let $R = \{R_1, \cdots, R_L\}$ be the set of translation references for all utterances. $R_l$ is defined as the reference set of the $l$-th utterance. In the experiments human translators paraphrased 16 reference sentences for each utterance. Therefore, $R_l$ contains 16 reference candidates.

$D(\hat{E}, R)$ is a translation “distortion” or an objective translation assessment. There are various proposed automatic evaluation metrics. However, BLEU and mWER metrics were used specifically in this study:

- BLEU [9]: A weighted geometric mean of the n-gram matches between test and reference sentences multiplied by a brevity penalty that penalizes short translation sentences.
- mWER [10]: Multiple reference word error rate, which computes the edit distance (minimum number of insertions, deletions, and substitutions) between test and reference sentences.

Because the objective function in the model (Eq. 3) is not a smoothed function, we used Powell’s search method to find a solution. Powell’s algorithm used in this work is similar to the one from [11] but we modified the line optimization codes, a subroutine of Powell’s algorithm, with reference to [3].

Finding a global optimum is usually difficult in a high dimensional vector space. To make sure that we find a good local optimum, we re-started the algorithm by using various initializations and used the best local optimum as the final solution.
Table 1: Comparisons of speech recognition performance with single-best and N-best hypotheses in terms of word accuracy, sentence accuracy, insertion, deletion and substitution error rates

<table>
<thead>
<tr>
<th></th>
<th>word acc(%)</th>
<th>sent acc(%)</th>
<th>ins (%)</th>
<th>del (%)</th>
<th>sub (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-best</td>
<td>93.5</td>
<td>78.7</td>
<td>2.0</td>
<td>0.8</td>
<td>3.6</td>
</tr>
<tr>
<td>N-best</td>
<td>96.1</td>
<td>87.0</td>
<td>1.2</td>
<td>0.3</td>
<td>2.2</td>
</tr>
</tbody>
</table>

different languages: Chinese, Japanese, Korean and English. Only Japanese-English parallel data was used in this study. The models used for translation were trained on acoustic models of our automatic speech recognition system, while the text database was used for training language and translation models.

The BTEC standard training set, containing 162,318 sentences, was used as the training data in this study while the standard test set #1 (510 sentences) and test set #2 (508 sentences) were used for development and test respectively.

The speech recognition engine used in the experiments is an HMM-based, large vocabulary continuous speech recognizer. The acoustic HMMs were context-dependent triphone models with 2,100 states in total, using 25 dimensional, short-time spectral features. In the first and second pass of decoding, a multiclass word bigram of a lexicon of 37,000 words plus 10,000 compound words was used. A word trigram was used for rescoring the recognition results.

The machine translation system is a graph-based decoder. The first pass of the decoder generates a word-graph, a compact representation of alternative translation candidates, using a beam search based on the scores of the lexicon and language models. In the second pass an $A^*$ search traverses the graph. The edges of the word-graph, or the phrase translation candidates, are generated by the list of word translations obtained from the inverted lexicon model. The phrase translations extracted from the Viterbi alignments of the training corpus also constitute the edges. Similarly, the edges are also created from dynamically extracted phrase translations from the bilingual sentences [13]. The decoder used IBM Model 4 with a trigram language model and 5-gram part-of-speech language models. The training of IBM Model 4 was implemented by the GIZA++ package [14].

To optimize the $\lambda$ parameters of log-linear models, we used the 510 sentences in the development data. For each utterance, $N \times K$ candidate translations were generated in the speech recognition and translation modules, where $N$ is the number of recognition hypotheses and $K$ is the number of translation hypotheses. A vector of $M=12$ feature values was attached to each candidate translation. We used the Powell’s algorithm described in the previous section to optimize these parameters. In the experiments, we set $N=100$ and $K=1,000$. We used a large $K$ to ensure that promising translation candidates were not pruned out. Experience has shown that this was indeed the case when the translation models were properly trained.

By using the different automatic translation evaluation metrics described in section 3, we obtained two sets of optimized parameters, corresponding to the BLEU and mWER metrics used, respectively.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>mWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-best</td>
<td>59.2</td>
<td>37.3</td>
</tr>
<tr>
<td>N-best without SR</td>
<td>61.1</td>
<td>37.6</td>
</tr>
<tr>
<td>N-best with SR</td>
<td>62.1</td>
<td>36.2</td>
</tr>
</tbody>
</table>

Table 2: Translation performance using the single-best speech recognition hypothesis and the N-best speech recognition hypotheses

4.2. N-best Hypothesis-based Translation

All 508 utterances in the test data were used to evaluate our proposed approach. Similar to processing the development data, the speech recognizer generated N-best ($N=100$) recognition hypotheses for each test speech utterance. Table 1 shows the speech recognition results of the single-best and N-best hypotheses of the test data set. We found that the N-best recognition hypotheses achieved more than 8% sentence accuracy improvement over the single-best hypotheses.

The speech recognition N-best output was then translated into corresponding English sentences. 1,000 such candidate translations were produced for each hypothesis. Finally a metric-specific log-linear model was used to rescoring these translation candidates and the one with the best score was selected. Table 2 shows the translation performance of each model.

In the table ‘single-best’ indicates translation of the single best hypotheses of speech recognition. ‘N-best without SR’ indicates the case of translations of N-best recognition hypotheses without using the speech recognition features in the log-linear model. ‘N-best with SR’ indicates the translations of N-best with the speech recognition features built into the log-linear model. The metrics in the rows are used for parameter optimization while the metrics in the columns are for evaluation.

In the experiments we found that the best translation was achieved when a relatively smaller set of recognition hypotheses, $N=5$, was used. Hence, the values in Table 2 were obtained by setting $N$ to 5.

The following observations can be made:

- The translations achieved by using the N-best speech recognition hypotheses were much better than those by the single-best. The cross validation with different translation metrics yielded consistent results. Note that for BLEU a high value means a better translation, but for mWER it means a worse translation.
- In most cases the best results were obtained by using the same metric as that used to optimize the $\lambda$ parameters.
- By comparing ‘N-best without SR’ with ‘N-best with SR’, we found that translation performance worsened when speech recognition features were not used.

4.3. Recognition Improvement for Incorrectly Recognized Sentences

In previous experiments we demonstrated that the speech translation performance was improved by exploiting N-best recognition hypotheses. In fact, the translation improvement for overall test data was the result of the large translation improvement for incorrectly recognized sentences because they tend to be the major cause of translation performance degradation.
Table 3: N-best translation improvement of incorrectly recognized utterances over the single-best

<table>
<thead>
<tr>
<th></th>
<th>BLEU (%)</th>
<th>mWER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-best</td>
<td>29.0</td>
<td>59.7</td>
</tr>
<tr>
<td>N-best</td>
<td>36.3</td>
<td>54.4</td>
</tr>
</tbody>
</table>

Table 4: Recognition accuracy of incorrectly recognized utterance improved by N-best hypothesis translation.

<table>
<thead>
<tr>
<th></th>
<th>word acc. (%)</th>
<th>sent. acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-best</td>
<td>74.6</td>
<td>0</td>
</tr>
<tr>
<td>N-best</td>
<td>76.4</td>
<td>7.5</td>
</tr>
<tr>
<td>mWER</td>
<td>75.9</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Here we carried out the following experiments. Only incorrectly recognized sentences were used for translation. The translation results are shown in Table 3. We found that the translation of incorrectly recognized sentences was improved significantly. The BLEU score was improved by 7.3% and mWER was improved by 5.3%.

Because we used N-best recognition hypotheses, the “Rescore” module chose the recognition hypothesis among the N hypotheses which yielded the best translation by the proposed models. As a consequence, speech recognition accuracy can be improved if the more accurate hypotheses were selected for translation. Table 4 shows the word accuracy and sentence accuracy of the recognition hypotheses selected by the metric-specific translation models. Comparing with the 74.6% word accuracy and 0% sentence accuracy of the baseline single-best speech recognition result, there was a significant improvement of speech recognition accuracy.

5. Conclusion

In this paper we have shown that N-best speech recognition hypotheses can improve speech translation over the single-best hypothesis. This was verified in the experiments by using a statistical machine translation system and constructing feature-based log-linear models. The robust SMT system achieved a significant performance improvement and the log-linear models were shown to be effective in incorporating multiple sources of features.

Using N-best recognition hypotheses in translation may be computationally costly when a large N is needed. However, our experiments have shown that a smaller N seems to be adequate to achieve most of the translation improvement. Moreover, a confidence measure approach in speech recognition [15] can be used to reduce computation further, where the confidence measure can be used to separate incorrectly recognized utterances from correctly recognized utterances. N-best hypothesis translation is then applied only to the utterances recognized with lower confidence while single-best translation is enough for utterances recognized with higher confidence. This can further improve the overall translation performance in terms of both speed and accuracy.

6. Acknowledgments

The authors would like to thank Dr. Seiichi Yamamoto and Dr. Yoshinori Sagisaka for their support. The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled “A study of speech dialogue translation technology based on a large corpus”.

7. References