Improved Speech Recognition Word Lattice Translation by Confidence Measure

Abdulvohid BOZAROV, Yoshinori SAGISAKA
Graduate School of Information and Telecommunication Studies
Waseda University, Tokyo, Japan
{abdulvohid@asagi.waseda.jp, sagisaka@giti.waseda.jp}

Ruiqiang ZHANG, Genichiro KIKUI
Spoken Language Translation Research Laboratories
ATR Laboratories International, Kyoto, Japan
{ruiqiang.zhang, genichiro.kikui} @atr.jp

Abstract

In conventional speech translation systems, Automatic Speech Recognition (ASR) produces a single hypothesis which is then translated by the SMT system. The translation results of SMT system are impaired by the word errors of the first best hypothesis in this approach more or less. To improve speech translation, we use a new word lattice translation approach which integrates multiple information sources from the speech recognition word lattice to discount the misrecognition. Furthermore, in order to improve speech translation and to reduce computation, we used N-bests cutoff, merging of identical word ids, and confidence measure. Experiments of Japanese-to-English speech translation showed that the proposed word lattice translation outperforms the conventional single best method.

1. Introduction

Due to the lack of robustness in automatic speech recognition (ASR), speech translation cannot achieve the same level of translation performance as achieved in text translation. The errors in speech recognition degrade the translation system performance to some extent. The tighter integration of the ASR and the machine translation (MT) is one of the ways to overcome this problem. Several architectures for speech translation have been proposed so far. For instance, Coupling of acoustical and translation models by statistical approach was tested by Ney [1]. Gao [2] employed a unified structure, where the maximum entropy approach was applied to integrate all features from ASR and MT together. A tightly connected system was proposed where finite-state transducer network is built to represent bilingual features by Casacuberta et al [3] [16].

Using speech recognition N-best hypotheses and integrating features from speech recognition and translation for speech-to-speech translation was proposed by Zhang et al [4]. Though this work effectively used N-best recognition hypotheses, it is only restricted to compensate the word errors of single best to improve speech translation. In this work we directly use a word lattice for speech translation. This approach has the same advantages as in [4] by keeping multiple ASR hypotheses as a word lattice in more compact, and computationally effective ways. In addition, speech recognition features, such as acoustic model, language model scores and confident measure, can be passed to MT component.

Using word lattice for speech translation was appeared recently in [5] where authors used CMU statistical machine translation (SMT) system [6] for text translation to carry out experiments on speech translation and observed translation improvement. In this paper we use sentence level posterior probability of recognition hypotheses [7] instead word level acoustic scores. We integrate all the features from ASR and SMT by using statistical log-linear model. New word lattice minimization approach is used to remove duplicated words and hypotheses. For evaluating our translation quality, we used widely used automatic evaluation criteria, including BLEU [8], NIST [9], mWER, and mPER.

2. The structure of proposed word lattice translation

The proposed speech translation structure is illustrated in Figure 1. It consists of two major parts: an automatic speech recognition (ASR) module and a speech recognition word lattice translation (SRWLT) module. The interface between these two components is speech recognition word lattice (SWL).

The task of speech translation for Japanese-to-English can be modeled as finding the target English sentence $\hat{E}$ which maximizes the probability $P(E|X)$, where X is a source Japanese utterance. If the intermediate output by ASR is defined as J, we get:

$$\begin{aligned}\hat{E} &= \arg\max_E P(E \mid X) = \arg\max_E P(E)P(X \mid E) \\ &= \arg\max_E \{P(E)\sum_J P(X, J \mid E)\} \\ &= \arg\max_E \{P(E)\sum_J P(X \mid J)P(J \mid E)\}\end{aligned} \quad (1)$$

where $P(J|X)$ is the acoustic model; $P(J|E)$ is the translation model; $P(E)$ is the target language model. We can approximate speech translation model described in Eq.(1) by dividing it into two steps:

- The ASR component generates word lattice G for the source language. Only hypotheses that have higher ASR scores than a threshold TH are kept in the word lattice.

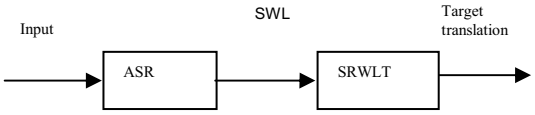

Figure 1: Structure of the proposed speech translation

$$G = \{J \mid P(J)P(X \mid J) > TH\} \quad (2)$$

- In the second step, SRWLT finds the output which maximizes:

$$< \hat{E}, \hat{J} >= \arg\max_{E,J} \{P(E)P(X \mid J)P(J \mid E)\} \quad (3)$$

where $J \in G$.

The speech translation structure shown in Fig.1 is an approximate implementation of the speech translation models, Eq. (1). In doing this, we assume J and E independent to derive Eq.(2) and approximate the summation by maximization to derive Eq.(3). As a side effect of the second step, a source $\hat{J}$ is also obtained aligned to $\hat{E}$, indicating $\hat{E}$ is translated from $\hat{J}$.

Feature based log-linear model was found very effective in speech translation ([10], [4]). Hence we adopt log –linear model, and we can formalize Eq. (3) as:

$$\begin{aligned}\hat{E} = \arg\max_{E} \{&\lambda_0 \log P_{pp}(J \mid X) + \lambda_1 \log P_{lm}(E) \\ &+ \lambda_2 \log P_{lm}(POS(E)) + \lambda_3 \log N(\Phi \mid E) + \\ &\lambda_4 \log P_0(null) + \lambda_5 \log T(J \mid E) + \lambda_6 \log D(E,J)\}\end{aligned} \quad (4)$$

We use seven features from acoustic model, language model, and translation models. They are:

(a) ASR hypothesis posterior probability $P_{pp}(J \mid X)$. We used the posterior probability instead of the acoustic model score since the latter has large dynamic range and difficult to normalize. The posterior probability is calculated as follows [7]:

$$P(J \mid X) = \frac{P(X \mid J)P(J)}{\sum_{J_i} P(X \mid J_i)P(J_i)} \quad (5)$$

(b) Word sequence target language model, $P_{lm}(E)$.

(c) Part-of-speech sequence target language models, $P_{lm}(POS(E))$.

(d) Fertility model, $N(\Phi \mid E)$. The probability of the English word, e, generating ϕ words.

(e) NULL translation model, $P_0(null)$. The probability of inserting a NULL word.

(f) Lexicon model, $\mathrm{T}(J \mid E)$. The probability of the word, j, in the Japanese source sentence being translated into the corresponding word, e, in the English target sentence.

(g) Distortion model, $\mathrm{D}(\mathrm{E},\mathrm{J})$. The alignment probability of the source and target sentence, (E,J).

Eq.(4) is a logarithm extension of eq.(1) where the translation model $P(J \mid E)$ is extended by IBM model 4 [11] and the acoustic feature is replaced by the posterior probability. The translation model was trained by GIZA++ [12].

3. Details for word lattice translation

Word lattice translation is much more complicated than text translation. In contrast to text translation where a single source is known, in lattice translation, there are multiple hypotheses to MT component. Which hypothesis is the best one to be translated is unknown before the decoding is completed. The statistical machine translation decoding is not time synchronous. The later part of the word lattice may be visited earlier than the front part. As the decoding proceeds, both the target sentence hypothesis and the source sentence hypothesis are updated based on Eq.(4). When the decoding is completed, the target sentence is found and aligned to a hypothesis in the source language.

To obtain the log-linear model feature weights in Eq.(4) for Japanese-to-English translation, we used two different approaches.

a) Direction set (Powells algorithm) method is one of the standard methods for multidimensional maximization and minimization problems. It is simple and doesn't require of gradient calculations [13].

b) We also used the approach by [10] by modifying to our problem (defined as Xpoints approach in Table 1). It is an algorithm for efficient line optimization using log-linear model, which is guaranteed to find the optimal solution.

As mentioned in [4], parameter optimization of log-linear models is important. We optimized log-linear models on different translation metrics by using each of these approaches. Some results of optimizing on BLEU metric by two of these approaches are shown in Table 1. First we set all the parameter values equal to 1, and get the translations. We can consider this translation result as a baseline, and compare translation results after parameter optimization with this result to see the improvements. The results for optimized parameters were obtained by setting hypotheses number N=20 and not using confidence measure. Both these two approaches produced better translations than that of the baseline. By comparing the results of these methods, we found that optimizing by both of two approaches significantly improved translation. Since Powells approach produced better results, we used the parameters obtained by this approach in the following up experiments.

Table 1: Optimization results.

	BLEU	NIST	mWER,%	mPER,%
Baseline	0.4324	6.2764	50.291	45.829
Powells	0.4569	6.7823	47.703	43.324
Xpoints	0.4479	6.4723	48.945	44.015

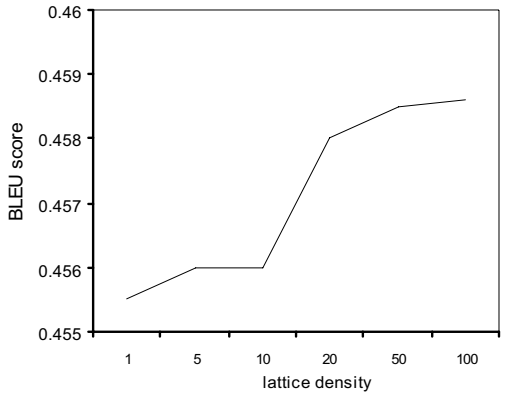

Figure 2:Changes in BLEU score by changing lattice density.

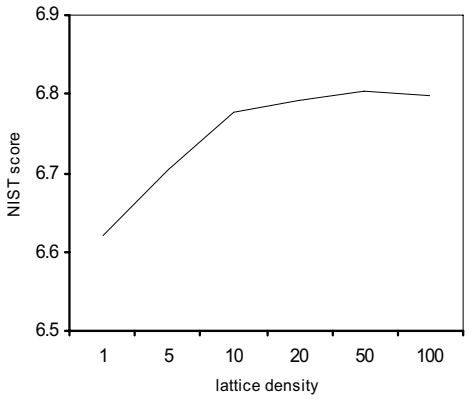

Figure 3: Changes in NIST score by changing lattice density.

We used the graph + A* decoding approach for the word lattice translation. It has been used for text translation in [15]. This is two pass decoding. The first pass uses a simple model to generate word graph to save the most likely hypotheses. It amounts to converting a source language word lattice (SWL) into a target language word graph (TWG). In the second pass, more complicated model is used to find best hypothesis by traversing the target word graph. For details see [17].

The SWL generated by ASR is a raw SWL, because the ASR uses time –synchronized decoding algorithm to produce the raw SWL, the same word identity can be recognized repeatedly in a slightly different frame span, appearing in multiple hypotheses with different edge identities. The direct conversion of SWL to TWG causes duplicated computations and graph explosions in TWG. We adopt the following methods to reduce the word lattice:

1) We cut off all the hypotheses except the top N-best. As the result raw SWL will be minimized. We use N=100 in our experiments.
2) We merge all edges with the same word ids into one edge id, so we can reduce lattice size by 50%.
3) We also found that using recognition confidence measure to filter low confidence hypotheses can improve translation quality.

Posterior probability has been used as a confidence measure in ASR in some instances [14]. We use posterior probability as a confidence measure in our experiments. By using Eq.(5), we calculate posterior probabilities of all hypotheses, and compare these probabilities with that of the first-best hypothesis, $P_{first-best}$. If this value is larger than $P_{first-best}/T$, T is the confidence threshold, then the hypothesis can be used in lattice translation. Thus, the number of hypotheses for lattice translation is determined by confidence measure filtering.

4. Data for speech translation experiments

We used the Basic Travel Expression Corpus (BTEC) for training, development and test in our experiments. BTEC contains travel related phrases, sentences, and dialogs. Currently it covers 4 languages: English, Chinese, Korean, and Italian. Each utterance has corresponding translations for multiple languages. In our experiments we use standard BTEC training data to train our models. The acoustic models used in the experiments are HMMs of triphone models with 2100 states in total, using 25 dimensional MFCC features. The acoustic models were trained using the trained data. The speech recognition engine was driven by a multiclass word bigram of a lexicon of total 47,000 words and word trigram for rescoring. The improvements in speech recognition error rate by using N-bests are shown in Table 2. All the LMs were trained using transcripts of training data. The training data contains 152,170 sentences, about 1 million 60 thousand words. The BTEC test data #1 was used for development data to train log-linear models. It has 510 sentences. And BTEC test data #2 was used as a test data. It has 508 sentences. And we downloaded evaluation tools from NIST's MT evaluation web-site[1]

5. Word lattice translation results

In the experiments the ASR system outputs the lattice and by using the lattice minimization approach described above we generate the minimized word lattice. The minimized SWL is created by setting the number of ASR hypotheses to 100. Then minimized SWL is translated by SRWLT translation system. Even though we used 100 for hypotheses numbers, this number can be changed in the program option for hypotheses numbers. After using confidence measure filtering, the number of hypotheses may be reduced again. Translation results are shown in Fig.2 , Fig.3 and in Table 2. Fig.2 shows the changes in BLEU score with the increased lattice density, and Fig.3 shows the changes in NIST score with the increased lattice density. All these results are obtained by setting confidence threshold to T=10 and parameters optimized by Powells method are used. We used these values because they produced best translations. We define lattice density as the number of hypotheses used in constructing the TWG from the minimized SWL.

Detailed results of our experiments are shown in Table 3. in the table, N stands for the lattice density. By comparing the results in the Table 1 and Table 3, we can see that the use of confidence measure improved translation accuracy. Through the analysis of these results, we can conclude that: (a) Overall translation improvements are observed over the single-best translation by lattice translation. (b) This improvement shows that the lattice translation can make use of more appropriate hypotheses for translation rather than the first-best hypotheses.

[1] http://www.nist.gov/speech/tests/mt/

Table 2: Speech recognition word accuracy

Number of N-bests	Word accuracy (%)
1	93.5
3	95.1
5	95.6
10	95.7
50	95.9
100	96.1

Table 3: Translation Results.

	NIST	BLEU	mWER,%	mPER,%
N=1	6.619	0.4555	48.1663	43.5365
N=5	6.7058	0.4560	48.0332	43.4754
N=10	6.7778	0.4560	47.7931	43.4531
N=20	6.7914	0.4580	47.6821	43.3021
N=50	6.8045	0.4585	47.7035	43.3293
N=100	6.7973	0.4586	47.5482	43.1868

6. Conclusions

This paper describes the state-of-the-art work about speech recognition word lattice translation. A novel word lattice decoding algorithm is implemented, where we use word graph search approach to construct target word graph and then, log-linear model based A* method. New lattice reduction algorithm is applied to reduce the size of raw word lattice. Even though all the experiments by using lattice translation produce better results than single-best translations, the best result is achieved by using speech recognition confidence measure. By choosing hypotheses on the posterior probability, speech translation results were improved to a new higher level. Our work of using the confidence measure is first attempt in speech translation, and we found it as a promising approach for speech translations, and it can solve the problem of instability of translation improvements by increasing lattice density [5].

7. Acknowledgements

This work was done in ATR laboratories International while the first of the authors was conducting short-term research in ATR. The author would like to thank ATR SLT members and researchers for their advice and discussions.

8. References

[1] Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. of ICASSP'1999*, volume 1, pages 517-520, Phoenix, AR, March.

[2] Yuqing Gao. 2003. Coupling vs. unifying: Modeling techniques for speech-to-speech translation. In *Proc. Of the EuroSpeech'2003*, pages 365-368, Geneva.

[3] Francisco Casacuberta, Enrique Vidal, and Juan M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. In *Proc. Of speech-to-speech translation workshop.* Pages 39-44,Philadelphia, PA, July.

[4] Ruiqiang ZHANG et al. 2004. Improved spoken language translation using n-best speech recognition hypotheses. *Proc. of the 8th International Conference on Spoken Language Processing.* Pages 1629—1632. Jeju, Korea.

[5] Shirin Saleem, Szu Chen Jou, Stephan Vogel, and Tanja Schulz. 2004. Using word lattice information for a tighter coupling in speech translation systems. In *Proc. of the ICSLP'2004*, Jeju, Korea.

[6] Stephan Vogel, Ying Zhang, Fei Huang, Alex Waibel. 2003. The CMU statistical machine translation system. In *Proc. of MT summit IX*, LA, USA.

[7] Lidia Mangu, Eric Brill, Andreas Stolcke. Finding consensus in speech recognition: word error rate minimization and other applications of confusion networks. *Computer Speech and Language*. Pages 373-400. (2000)14.

[8] Papineni Kishore, Roukos Salim et al. 2002. BLEU: A method for Automatic Evaluation of Machine translation. *Proc. Of the 20th Annual Meeting of the Association for Computational Linguistics*.

[9] NIST report. 2002. Automatic Evaluation of Machine translation quality using n-gram co-occurrence statistics. *http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.*

[10] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL'2003*, pages 160-167.

[11] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263-311.

[12] Franz Josef Och, Herman Ney. Improved Statistical Alignment Models. *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hong Kong, China, October 2000.

[13] William H. Press, Saul A. Teukolsky, William T. Vetterlng, and Brian P. Flannery. 2002. *Numerical recipes in C++*. Cambridge University Press, Cambridge, UK.

[14] Frank Soong, Wai-Kit Lo, and Satoshi Nakamura. 2004. Generalized word posterior probability for measuring reliability of recognized words. In *Proc. of the 2004 Special Workshop in Maui*, Hawaii.

[15] Nicola Ueffing, Franz Josef Och, and Hermann Ney. 2002. Generalized word graphs in statistical machine translation. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP02)*, pages 156-163, Philadelphia, PA, July.

[16] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, C. Martinez D. Llorens, S. Molau, F. Nevado, M. Pastor, D. Pico, and A. Sanchis. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25-47, 2004.

[17] Ruiqiang Zhang, Genchiro Kikui. Integration of speech recognition and machine translation: speech recognition word lattice translation. coming soon in Journal of *Speech Communication*. 2005