CONSOLIDATION BASED SPEECH TRANSLATION

Chiori Hori[†] *Bing Zhao*[‡] *Stephan Vogel*[‡] *Alex Waibel*[‡]

†NiCT-ATR 2-2-2 Hikaridai, Soraku-gun, Kyoto, 619-0288 Japan
‡Carnegie Mellon University, 407 S. Craig St., Pittsburgh, PA 15213
E-mail: †chiori-hori@atr.jp, ‡{bzhao, vogel, waibel}@cs.cmu.edu

ABSTRACT

To alleviate the degradation of the performance of speech translation, this paper proposes a new approach to translate ASR results through consolidation which extracts meaningful phrases and remove redundant and irrelevant information caused by speaker's disfluency and recognition errors. The speech translation results via consolidation are partial translation and can not be directly compared with gold standards in which all words are translated. We would like to propose a new evaluation framework for partial translation by comparing with the most similar set of words extracted from a word network created by merging ***gradual summarizations*** of the gold standard translation. Chinese broadcast news speech in RT04 were recognized, consolidated and then translated. The performance of MT results was evaluated using BLEU. We propose ***Information Preservation Accuracy*** (***IPAccy***) and ***Meaning Preservation Accuracy*** (***MPAccy***) for consolidation and consolidation-based MT.

Index Terms— Speech Consolidation, Machine translation, Chinese broadcast news speech, Chinese-English translation

1. INTRODUCTION

Speech translation systems are designed as systems to combine machine translation (MT) systems with automatic speech recognition (ASR) technology. Recognizing spontaneous speech with high accuracy remains a challenge for an ASR system. Text translation is still difficult and speech translation introduces additional difficulties caused not only by disfluencies due to the spontaneity of the spoken language but also due to the errors in the ASR output. Recognition errors, in particular, could potentially change the meaning of translated sentences. To accomplish more accurate speech translation, we have to solve problems in translating ill-formed spoken language and handling errorful ASR output. To avoid degradation of MT performance by speech recognition errors, we have proposed speech consolidation to get more reliable phrases which preserve the original meaning and contribute positively to the total performance of spoken language systems such as Summarization, MT and question answering (QA) [1]. We confirmed the consolidation approach extracted more reliable phrases from the ASR results for the Translanguage English Database (TED) corpus by comparing with the manual consolidation results[1]. In the next step, we need to examine how the consolidation enhances the performance of language post processing.

This paper presents a speech translation system in which ASR, consolidation and MT systems are cascaded. Mandarin news speech in RT04 was recognized, consolidated, and translated into English text. We evaluated both intrinsic and extrinsic performance of the speech consolidation i.e., consolidation accuracy and MT performance effected by consolidation, respectively.

2. CONSOLIDATION-BASED SPEECH TRANSLATION

Our system is designed as ASR [2], consolidation (CON) and statistical MT (SMT) systems [3] are cascaded in Fig. 1.

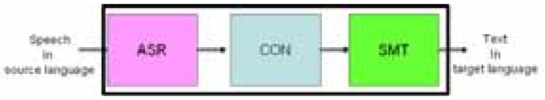

Figure 1. Consolidation-based Speech Translation System

We propose a speech consolidation approach using confusion networks obtained by ASR systems [1]. Figure 2 shows an example of consolidation of a confusion network. The consolidation result is "<s> today I would like </s>" from the confusion network.

[1] TED (http://www.elda.org/catalogue/en/speech/S0031.html)

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

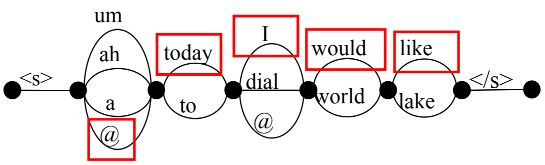

Figure 2. Example of consolidation of confusion network

Our approach for speech consolidation is based on summarization by word extraction [4]. This summarization approach extracts a set of words from ASR output by focusing on 1) extracting informative phrases, 2) excluding redundant and irrelevant phrases, and 3) concatenating extracted words in summarization to retain the original meaning according to a given compression ratio using a Dynamic Programming (DP) technique. We split the summarization approach into two phases i.e., removing garbles and extracting salient information. The consolidation handles the former function by combining 2) and 3). The consolidation technique attempts to extract a much longer phrase preserving the original meaning. The compression ratio is not given to the consolidation process. This makes consolidation more difficult than the summarization with a compression ratio. To solve this problem, we've incorporated the skip and insertion penalties used in ASR approach into the consolidation in Eq. (1). We select a set of words which maximize the following consolidation score from all possible combinations of words in an utterance using DP.

$$S(V)=\sum_{m=1}^{M}\{\lambda_L L(v_m \mid v_1 \ldots v_{m-1})+\lambda_C C(v_m)+sp\cdot d(v_{m-1},v_m)+ip\} \quad (1)$$

sp : a skip penalty (sp<0)
$d(v_{m-1},v_m)$: the number of skipped words between v_{m-1} and v_m
ip : a insertion penalty
$L(v_m|v_1,\ldots,v_{m-1})$: linguistic score $P(v_m|v_{m-2},v_{m-1})/P(v_m)$
$C(v_m)$: confidence score

The *sp* is incorporated to avoid connecting two words located a long distance apart in the original sentence. We proposed a dependency score to avoid concatenating two words which do not have dependency structure [4]. Dependency detection in spontaneous speech is still challenging and we use the skip penalty instead of the dependency score in this study.

The *ip* is used to control the total compression ratio to avoid consolidating speech with high compression, since high compression of a sentence often alters the meaning of the sentence. The skip and insertion penalty is determined experimentally.

3. EVALUATION

We evaluated both intrinsic and extrinsic performance of the speech consolidation i.e., consolidation accuracy and MT performance effected by consolidation, respectively. This paper proposes a new evaluation framework for consolidation and partial translation by comparing with the most similar set of words extracted from a word network created by merging ***gradual summarizations*** of the gold standard transcription and translation.

3.1. Intrinsic evaluation

3.1.1. Gold standards for automatic consolidation

The gold standards for the ASR results were prepared by humans. Recognition errors can be changed under different recognition conditions and thus we need a gold standard for each recognition result. It is too hard to get such manual consolidation results for each recognition result. To alleviate this labor, this paper proposes an evaluation method using gradual summarization networks.

The goal of consolidation is to extract meaningful phrases which preserve part of the original meaning from ASR outputs by removing recognition errors and fragments. In the first step, we delete substitution and insertion errors from ASR results by comparing with manual transcriptions. Supposing an utterance (TRS) is misrecognized as in Table 1, we delete "boot", "full" and "chill" to generate the ASR results excluding errors automatically. We denote the word string without ASR errors as "**DEL**".

Table 1. Example of ASR and consolidation results
TRS: transcription, DEL: ASR results excluding errors

TRS	<s> The beautiful cherry blossoms in Japan bloom in spring </s>
ASR	<s> The (beat) (full) (chill) blossoms in Japan bloom in _____</s>
DEL	<s> **The** ___________ blossoms in Japan bloom **in** _____ </s>
CON	<s> __________ blossoms in Japan bloom __ ____ </s>

In the second step, we have to delete some fragments which are correctly recognized, i.e. "The", "in" in the **DEL** row. Humans can delete such fragments to get consolidation (CON) by judging isolated words using context. While we need to know which set of words retain the original meaning to clean up such fragments automatically.

When the dependency structure of the TRS is known, it is not difficult to detect fragments. The dependency structure of spoken language is still difficult to be estimated accurately. In this evaluation, we use gradual summarization networks to detect original dependency. The original utterance (TRS) in Table 1 is manually summarized by deleting words step by step as shown in Table 2. This gradual summarization process is almost the similar process by "Parsing by Chunks" [5].

Table 2. Example of gradual summarization
ORG: original speech, GRDSUM: gradual summarization

ORG	<s> The beautiful cherry blossoms in Japan bloom in spring </s>
GRD SUM	<s> The _______ cherry blossoms in Japan bloom in spring </s>
	<s> The _______ cherry blossoms __ ____ bloom in spring </s>
	<s> The_______ cherry blossoms __ ____ ______ in spring </s>
	<s> The _______ cherry blossoms __ ____ _____ __ ______ </s>
	<s> ___ _______ _____ blossoms __ ____ _____ __ ______ </s>

These gradual summarizations are merged into a word network in Fig. 3. This network represents dependency structure between words. All phrases from <s> to </s> extracted from the word network retain part of the original meaning.

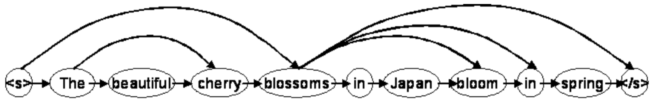

Figure 3. Example of a gradual summarization network

The **DEL** consisting of only correctly recognized words in the ASR output, “<s> The blossoms in Japan bloom in </s>” is compared with the above gradual summarization network. A set of words minimizing errors based on Levenshtein distance is extracted. However, the consolidation cannot use words which are misrecognized in the ASR results in the gradual summarization network. In this example, “*beautiful*”, “*cherry*” and “*spring*” are not available when consolidating. We select a word string which minimizes errors among the word strings consisting of the correctly recognized words in the gradual summarization network and used it as a gold standard for consolidation.

We note that the gradual summarization network does not cover all word strings which retain part of the original meaning. However, these missing word strings which preserve the original meaning do not give a favor to automatic consolidation. Indeed they are counted as errors of automatic consolidation because correct word concatenations which are not covered by the gradual summarizations are not counted as correct answers.

3.1.2. Evaluation metric

We define two types of measures based on word accuracy. One is for meaning preservation and the other is for information preservation. We denote ***meaning preservation accuracy*** and **information preservation accuracy** as ***MPAccy*** and ***IPAccy***, respectively.

MPAccy shows to what degree the consolidation results preserve the original meaning. The degree of preservation of the original meaning via consolidation is evaluated by comparing a consolidation result and an extracted gold standard from a **gradual summarization network.** Since the gradual summarization network is aimed to be gold standards for part of phrases with various lengths, the consolidation has the potential to achieve 100% ***MPAccy***.

IPAccy shows how much information in the original utterance is preserved. The compression ratio of the gold standards shows the upper bound of information preservation by consolidation. Since ***MPAccy*** shows what degree of information in the gold standard is preserved by consolidation, we can see the total performance against the original utterance using Eq. (2).

$$\boldsymbol{IPAccy} = \boldsymbol{MPAccy} * \text{CR(Gold Standard)} \quad (2)$$

where CR shows compression ratio of the extracted gold standard. Our consolidation approach cannot preserve the same amount of information in original speech due to recognition errors and thus ***IPAccy*** has an upper bound.

3.2. Extrinsic evaluation

The whole cascaded system of ASR, consolidation and MT systems is evaluated in terms of how much consolidation can enhance the performance of speech translation. We attempt to evaluate the performance of MT via consolidation. Since consolidation results missed some phrases in speech, we cannot evaluate the real contribution by consolidation when comparing with references in which all words in source speech are translated. To evaluate partial translation based on consolidation, we need manual translations for each partial transcription.

3.2.1. Synchronous gradual summarization

The ideal gold standards for consolidation-based speech translation are manual translations of all “part of phrases” which retain part of original meaning in speech. The bilingual translator generates one manual reference in response to each utterance in the source side and then gradually summarizes both the source and target sides by extracting words synchronously. Table A in Appendix shows an example of a process to generate ***synchronous gradual summarizations***. The transcription and its’ translation are gradually summarized according to 6 steps. Each set of gradual summarizations of both sides has the same meaning. We get multiple manual translations with various lengths. To cover more word strings which preserve part of the original meaning in the manual translations, the gradual summarizations are merged into a word network as described in the section 3.1.

3.2.2. Reference and metrics

We prepare three types of references for MT, i.e., one reference which is a manual translation for each manual transcription and multiple references which are manual translations for all gradual summarizations. To prepare references for consolidation-based MT, the gradual summarization in the target side are merged into a word network. A set of words maximizing word accuracy is extracted from the gradual summarization network by comparing with consolidation-based MT. The length of the extracted word string is almost the same as that of the consolidation-based MT. We set the extracted word string as a certain reference in response to the consolidated-base MT. We compare the performance of the consolidation-based MT results with the gold standards using BLEU and ***MPAccy***.

4. EXPERIMENT

4.1. Data

4.1.1 Speech data and manual transcription

297 utterances (627 sentences) of Chinese BN speech in RT04 were recognized and then translated into English text. To test the performance of consolidation, we used 125 utterances from the beginning of RT04 were consolidated and translated. The manual transcriptions for all speech are provided by LDC[2].

4.1.2. Synchronous gradual summarization

To generate gold standards for consolidation and consolidation-based MT, the bilingual translator translated the 125 Chinese manual transcriptions into English text and then gradually summarizes both Chinese and English sides by removing words synchronously. The Chinese gradual summarization was used in the intrinsic evaluation for the consolidation accuracy and the English gradual summarization was used in the extrinsic evaluation for the consolidation-based translation.

4.2. ASR system

The ISL RT04 Mandarin Broadcast News evaluation system using the JANUS speech recognition toolkit was applied to the speech translation system [2]. The acoustic models were trained using 27 hours of the Mandarin HUB4 1997 training set and 69 hours of the TDT4 Mandarin data. 42-dimension features after Linear Discriminant Analysis were used for the front-end processing. The system employs a multi-pass decoding strategy in which cross adaptation among the syllable-based and the phone-based decoders were performed. The vocabulary size is 63K word. Confusion word networks were given to the consolidation system.

The test set was RT04, consisting of 297 utterances segmented for evaluating ASR system. Speech data involves English speech other than Mandarin speech and thus some speakers were entirely misrecognized. In addition, some of the speech data were dialogues with a very spontaneous style. The character and word errors of our ASR system were 21.2% and 46.8%, respectively. The word accuracy was affected by mismatches of word segmentation between ASR output and manual transcription.

4.3. Statistical Machine Translation system

Manual transcription of speech and recognition results were translated using the CMU SMT system based on phrase-to-phrase translations [3]. The experimental conditions were described in the following section.

4.3.1. Phrase alignment model

760471 sentence pairs were sub-sampled for the TIDES '02, '03 and '04 test sets from a 200 million words parallel corpus. The feature of data is listed in Table 4. Phrase table contains 1666428 entries ranging from 1-gram to 10-gram on the source side. There are eight score functions for each phrase pair [6].

Table 4. Data used for phrase alignment model.

	#sentences	#words	#characters
Trains04.split.en	760471	12524365	72357758
Train04.split.gb	760471	11484629	46091860

4.3.2. Baseline performance using Chinese newspaper text

The SMT system was constructed for translating Chinese newspaper text. The test sets provided in TIDES '02 can be translated with BLEU=27.22 (length penalty=1) and NIST= 8.7143 (length penalty = 0.9942).

4.3.3. MT Performance for ASR output

Speech recognition output with 21.2 % character error rate was translated with BLEU=8.20 and NIST=4.1425. The difference between translating the manual transcription and the ASR output is not significant. The results show that the degradation of the performance of translating BN against translating newspaper text is mainly caused mismatch features in the model between BN and newspaper text. Table 5 lists out-of-vocabulary (OOV) rate and perplexity (PP) for each test sets.

Table 5. Out-of-vocabulary rate and perplexity

			#sentences	#words	OOV	PP
TIDES ‘02	Source		878	24337	0	138
	4 references		3512	105143	0	148
RT04	Source	TRS	297	11547	0	536
		ASR	297	9724	0	848
	1 reference		297	12105	0	300

The OOV and PP for the source side was calculated using the trigram of the source text in the parallel corpus for the

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S16

alignment model and those for the target side was calculated using the trigram used in the SMT decoder. There is a big mismatch between the training and test data in the RT04.

5. EVALUATION RESULT

Figure 4 shows the results of the intrinsic evaluation based on character. The manual transcription, the ASR results, and the consolidation with and without language model score were evaluated.

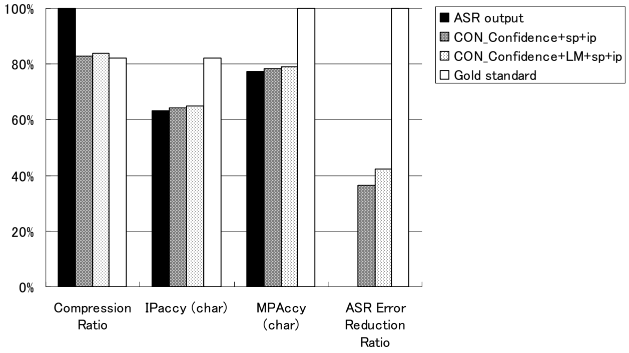

Figure 4. Performance of automatic consolidation

The compression ratio of the consolidation is almost the same as the gold standard. The consolidation approach worked well to detect the length automatically. While ***MPAccy*** based on character of the ASR results was 77.2%, our consolidation approach achieved 79.2% ***MPAccy***. To evaluate how many recognition errors are removed from ASR results, the reduction ratio of speech recognition errors were calculated. 42.4% of insertion and substitution errors in the ASR output were removed by consolidation. Our consolidation achieved a higher ***MPAccy*** and a higher reduction ratio of recognition errors than those for the ASR output. These results show that our consolidation approach can extract a set of phrases which preserve the original meaning by excluding fragments and recognition errors.

Figure 5 shows the extrinsic evaluation for consolidation-based MT. The MT of the manual transcription, the ASR results, and the consolidation with and without language model score were evaluated. The lengths of the MT of the manual transcription, the ASR output, and the consolidation result were 74%, 67%, and 60% of the manual translation of the manual transcription, respectively.

We evaluated the MT based on BLEU (N=4) using one reference. There was no difference among the MT with and without consolidation. The BLEU using all gradual summarizations with different lengths shows that the results for ASR, CON and Gold standards are 10.04, 11.00 and 12.09 BLEU scores, respectively. This results show the consolidation contributed to enhance the MT performance.

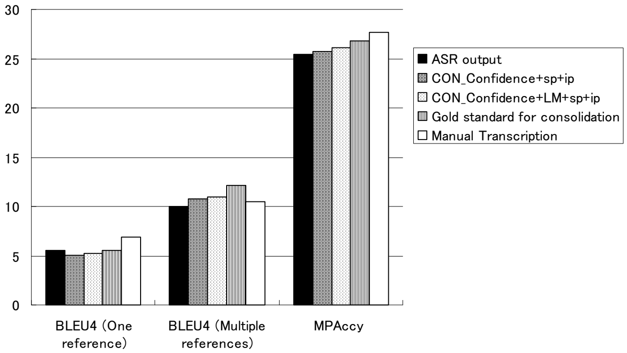

Figure 5. Performance of machine translation

However, the performance for the MT of the manual transcriptions is lower than others and was not evaluated well here. We discuss reasons for this phenomena in the below section for discussion.

MPAccy using a certain reference for the MT extracted from the **gradual summarization network** shows the MT performance was enhanced by the consolidation. In addition, ***MPAccy*** for manual translation is higher than others and it is appropriately evaluated.

6. DISCUSSION

Figure 4 shows that the consolidation preserved a part of meaning of the original speech whilst excluding the errors, but the BLEU using one reference does not show the difference in the MT performance between the ASR output and the gold standard for consolidation in Fig. 5. This is because the partial translations are not counted as correct translation even if the partial translation preserves part of the original meaning.

When we used the gradual summarizations as multiple references, the BLEU was increased by the consolidation. However, the score for the manual transcription was lower than those of others. There is a problem in the evaluation based on BLEU using multiple references with various lengths. The precision-based BLEU for shorter translations tend to be higher than that of longer translations. To give a penalty to precisions for shorter translations, BLEU is penalized by length of references when hypotheses are shorter than references. When we considered multiple references, the penalty is determined by the length of the reference which is the most similar to the MT result. All translations can find references which have the same length or similar length among the multiple references with various lengths and thus the penalty for length given to BLEU does not work well. The results just show that the "longer" translations resulted in "lower" precision. It is the reason why the MT results for the manual transcription is longer than others and shows the lower precision.

On the other hand, the translations of the manual transcriptions were fairly evaluated by ***MPAccy*** using the gradual summarization network. The results of ***MPAccy*** show that the consolidation enhanced the performance of MT.

7. CONCLUSION

This paper proposed speech consolidation-based translation. The RT04 Mandarin Broadcast News speech was recognized, consolidated and translated into English text. We confirmed that automatic consolidation extracts useful phrases effectively from the speech recognition results and the MT performance is enhanced by the consolidation. ***MPAccy*** using multiple references consisting of gradual summarizations of manual translations is capable of evaluating consolidation-based MT reasonably.

Future works will involve evaluating consolidation-based MT performance in response to various levels of speech recognition performance. Consolidation could alleviate the degradation especially when the speech recognition performance is low. Furthermore, consolidation can apply to translation results as well. We can compare the performance for consolidation before/after translation. Recently, speech translation is done on confusion networks obtained by ASR systems since the confusion network is compact and capable to keep multiple hypotheses. We can integrate consolidation directly into MT systems that translate from confusion networks.

8. REFERENCE

[1] C. Hori and A. Waibel, "Spontaneous Speech Consolidation for Spoken Language Applications," Proc. Interspeech2005.

[2] H. Yu, Y. Tam, T. Schaaf, S. Stueker, Q. Jin, M. Noamany and T. Schultz, ``The ISL RT04 mandarin broadcast news evaluation system," in EARS Rich Transcription Workshop, 2004.

[3] S. Vogel, "PESA: Phrase pair extraction as sentence splitting," in Proc. of the MT Summit X, 2005.

[4] C. Hori, S. Furui, R. Malkin, H. Yu and Al. Waibel, "A Statistical Approach for Automatic Speech Summarization," EURASIP Journal on Applied Signal Processing 2003:2, pp. 128-139.

[5] S. Abney, "Parsing by Chunks", "Principle-Based Parsing: Computation and Psycholinguistics", Kluwer Academic Publishers, 1991

[6] B. Zhao and S. Vogel, "A generalized alignment free phrase extraction," in Proceedings of the ACL Workshop on Building and Using Parallel Texts, 2005, pp. 141-144.

Appendix

Table A. Example of synchronous gradual summarization

	Chinese Transcription	English Translation
Original	第十五届中美商贸联委会举行双方签署八项协议和换文。	The Fifteenth China-US Commerce and Trade Coordinative Commission was held with eight agreements signed and notes exchanged by the two parties.
Step 1	________中美商贸联委会举行双方签署八项协议和换文。	The ______ China-US Commerce and Trade Coordinative Commission was held with eight agreements signed and notes exchanged by the two parties.
Step 2	________中美商贸联委会举行双方签署____协议和换文。	The _______ China-US Commerce and Trade Coordinative Commission was held with ____ agreements signed and notes exchanged by the two parties.
Step 3	________中美商贸联委会举行____签署____协议和换文。	The _______ China-US Commerce and Trade Coordinative Commission was held with ____ agreements signed and notes exchanged______________.
Step 4	________中美商贸联委会举行____签署____协议______。	The _______ China-US Commerce and Trade Coordinative Commission was held with _____ agreements signed ___ ____________________________.
Step 5	________中美商贸联委会举行____________________。	The _______ China-US Commerce and Trade Coordinative Commission was held with _____ agreements signed___ ____________________________.
Step6	________中美商贸联委会________________________。	The _______ China-US Commerce and Trade Coordinative Commission ______________________________________ ____________________________.