Spontaneous Speech Consolidation for Spoken Language Applications

Chiori Hori, Alex Waibel

InterACT, Language Institute of Technology
Carnegie Mellon University
{chiori,ahw}@cs.cmu.edu

10.21437/Interspeech.2005-65

Abstract

This paper describes the work done as a part of the International Workshop on Speech Summarization for Information Extraction and Machine Translation (IWSpS) [1], on spoken language processing including summarization, machine translation and question answering on lecture speech in the Translanguage English Database (TED) corpus [2]. The hypotheses of lecture speech obtained by automatic speech recognition (ASR) system are ill-formed due to the spontaneity of speakers and recognition errors. The overall performance of spoken language processing components is affected by the errors introduced by the ASR system. In order to get more reliable phrases which maintain the original meaning and contribute positively to the total performance of the spoken language system, this paper proposes a ***consolidation*** framework. The consolidation approach extracts words by excluding redundant and irrelevant information and concatenating words so as to maintain the original meaning. Automatic consolidation performance is evaluated by comparing with manual consolidation by humans using a word accuracy metric. Our approach gives 58% accuracy on ASR output with 70% word accuracy.

1. Introduction

In the past, we have worked on summarization of speech in meetings [1] and broadcast news [2] and machine translation (MT) of travel conversations in the C-star project [3], appointment negotiation in the Verbmobil project [4], and dialogue in e-commerce in the NESPOLE! project[5]. Currently, we are working on domain unrestricted speech translation tasks such as telephone conversations, lectures, meetings and broadcast news speech in the STR-DUST (Speech Translation for Domain-Unlimited Spontaneous Communication Tasks) project[6].

A spoken language processing system needs to be combined with language processing and automatic speech recognition (ASR) technologies. Summarization, Machine Translation (MT) and question answering (QA) on written text with large vocabulary such as newspaper text and HTML documents are being actively investigated using statistical approaches[7]. Such technologies are incorporated into speech processing. However, written text is still difficult even if huge corpora are available for calculating statistic models and speech processing is more complicated. The difficulty in speech processing is mainly caused by the style of spoken language which is different from written text. Spontaneous speech includes colloquial expressions and ill-formed sentences caused by spontaneous aspects such as incorrect grammar, incomplete sentences, and redundant expressions i.e., disfluencies, repetitions, word fragments. In addition, ASR output is not always perfect and we also have to handle recognition errors.

Recently, spontaneous speech recognition has been intensively investigated. English academic presentation speech was recognized by adapting models of written text to spoken language transcriptions [3] [4]. To detect phenomena in spoken language statistically, we need to collect spontaneous speech. Japanese academic presentation speech and free talk with various topics are manually transcribed and annotated precisely [5] and English broadcast news and conversational telephone speech are annotated with markers such as edit words in the EARS project[8].

Recognizing spontaneous speech with high accuracy remains a challenge for an ASR system. When we combine ASR and language processing, the total performance is affected by ill-formed sentences and incorrect information which is introduced by ASR system. This paper proposes ***consolidation*** framework to get more reliable phrases which maintain original meaning and contributes to the total performance of spoken language systems.

To consolidate transcription of speech, redundant information caused by disfluencies and irrelevant information by recognition errors should be deleted. Recognition errors are induced by disfluencies and OOV words. To handle disfluency, such as fillers, repetitions, corrections and false starts, we are working on disfluency removal [6]. Focusing on deleting OOV, OOV words are forcibly recognized as a word in the ASR vocabulary and not only the OOV word itself but words surrounded it are also accidentally misrecognized. Detecting OOV words is difficult in domain unrestricted task. On the other hands, confidence measures [7] can be applied to delete acoustically and linguistically unreliable phrases. However, the meaning of the phrase after deleting unreliable words sometimes does not correspond to the original meaning intended by the speaker.

To extract more reliable phrases which maintain original meanings, summarization by extracting words can be applied [2]. This approach extracts words from transcriptions according to compression ratio by focusing on 1) extracting important content words, 2) excluding redundant and irrelevant phrases, and 3) concatenating words in summarization to maintain original meanings. It accomplishes

[1] IWSpS (http://www.is.cs.cmu.edu/iwsps2004/)
[2] TED corpus (http://www.elda.org/catalogue/en/speech/S0031.html)
[3] C-STAR project (http://www.c-star.org/)
[4] Verbmobile (http://verbmobil.dfki.de/)
[5] NESPOLE! Poject (http://nespole.itc.it/)
[6] STR-DUST (http://www.is.cs.cmu.edu/str-dust/)
[7] TIDES project (http://www.ldc.upenn.edu/TIDES/)
[8] EARS project (http://www.nist.gov/speech/tests/rt/)

speech consolidation and important information extraction simultaneously. In this paper, we extend consolidation aspects in summarization. The consolidation approach proposed here attempts to extract not only important phrases but also all phrases which maintain original meanings without being given compression ratio.

2. Speech Consolidation

2.1. Speech Summarization for Consolidation

We proposed summarization framework by word extraction [2]. The summarization score indicating the appropriateness of a summarized sentence is defined as the sum of the linguistic score L of the word string in the summarized sentence, the word significance score I, the confidence score C of each word in the original sentence and the word concatenation score Tr. The word concatenation score given by SDCFG indicates a word concatenation probability determined by a dependency structure in an original sentence. This method is effective in reducing the number of words by removing redundant and irrelevant information without losing relatively important information. A set of words maximizing the total score is extracted using a Dynamic Programming (DP) technique.

Given a transcription result consisting of N words, $W=w_1,w_2,\ldots,w_N$, the consolidation is performed by extracting a set of $M(M < N)$ words, $V=v_1,v_2,\ldots,v_M$, which maximizes the score given by eq.(1).

$$S(V)=\sum_{m=1}^{M}\lambda_L L(v_m \mid v_1 \ldots v_{m-1})+\lambda_I I(v_m)+\lambda_C C(v_m)+\lambda_T Tr(v_{m-1},v_m), \quad (1)$$

where λ_L, λ_I, λ_C and λ_T are weighting factors for balancing among L, I, C, and Tr.

In consolidation, removing recognition errors retaining as much information of the original sentence as possible and reconstructing a fluent sentence are important factors. We modify the summarization score to function for effective consolidation as

$$S(V)=\sum_{m=1}^{M}\{\lambda_L L(v_m \mid v_1 \ldots v_{m-1})+\lambda_C C(v_m)+sp\cdot d(v_{m-1},v_m)+ip\}, \quad (2)$$

where sp is a skip penalty ($sp<0$); $d(v_{m-1},v_m)$ is the number of skipped words between v_{m-1} and v_m; ip is a insertion penalty. The skip penalty is incorporated to avoid high compression of the original sentence (i.e. low summarization ratio) because high compression of a sentence often alters the meaning of the sentence. ip is used to control the total summarization ratio.

The linguistic score $L(v_m|v_1,\ldots,v_{m-1})$ indicates the appropriateness of the word strings in a summarized sentence. It is measured by the logarithmic value of a trigram probability $P(v_m|v_{m-2},v_{m-1})$. For consolidation, since we focus only on connectivity between words, we use an adjusted trigram probability $P(v_m|v_{m-2},v_{m-1})/P(v_m)$ instead of the regular trigram. This normalized trigram removes the influence of frequency and represents only word concatenation correctness.

The confidence score $C(v_m)$ is incorporated in the above equation to weight acoustically as well as linguistically reliable hypotheses. Specifically, a posterior probability of each transcribed word, that is the ratio of a word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by a decoder and used as the confidence measure. In this study, we use a confusion network [8] instead of a word graph since more accurate posterior probabilities are derived from confusion network.

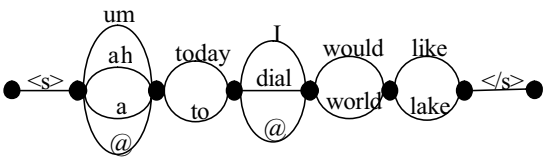

Fig.1 Confusion Network.

2.2. Consolidation Algorithm for Confusion Networks

A DP technique for speech summarization can directly be applied to speech consolidation. However, the algorithm is only for a 1-best hypothesis in speech recognition. In this work, we extend the algorithm so as to find the best consolidation result from among multiple hypotheses represented in a confusion network. The extended algorithm has a potential to reduce recognition errors by reselecting the words in the network through the consolidation.

A confusion network is a compact representation of multiple hypotheses generated in speech recognition. Figure 1 shows an example of the confusion network. Compared to a word lattice/graph, it is more compact since it ignores the connectivity of adjacent words and discards time information of each word. We assume that all sentences included in a confusion network begin with "<s>" and end with "</s>". Let N be the length of the confusion network, i.e. the number of confusion sets. The confusion set consists of a set of competing words in one column as in Fig.1. For example, the first confusion set includes only "<s>", and the second set includes "um", "ah", "a", and "@". The symbol "@" is a special word indicating a possibility of deletion. In a confusion network, a posterior probability is attached to every word. Sum of the probabilities in each confusion set becomes 1.

First we define a notation used in the algorithm:

f,g,h: a partial consolidated sentence hypothesis that has members of the score (*score*), the word sequence (*words*), and the position of the confusion set that the last word of the hypothesis is included (*pos*).

F,G,H,H' : hypothesis list that contains hypotheses,

$\hat{H}$: hypothesis list that contains complete hypotheses,

$\hat{h}$: the best consolidated sentence hypothesis,

Generate(): function that generates a new hypothesis,

Insert(*H*, *h*): function that inserts *h* into *H*,

Move(*H, F*): function that moves all hypotheses in *F* to *H,*

ExpandHypo(*h*): function that generates a list of new hypotheses by adding each word that can succeed *h*,.

CFNet(*n*): function that returns the *n*-th confusion set of words in the confusion network.

Second we describe the main procedure of the algorithm:

```
// Main procedure
begin
  h := Generate( )
  h.words := "<s>"
  h.pos := 1
  h.score := 0
```

```
  Insert(H, h)
  while H is not empty do begin
    foreach h ∈ H do begin
      F := ExpandHypo(h)
      foreach f ∈ F do begin
        if f.pos = N then  // Is f a complete hypo?
          Insert(Ĥ, f)
        else
          Insert(H', f)
      end
    end
    H := H'
    H' := ϕ        // clear all hypotheses in H'
  end
  ĥ := max_{h∈Ĥ} h.score
end
```

$\hat{h}.words$ is the most likely consolidation result. For simplification, a pruning step is omitted in the above description.

Finally we show the procedure of *ExpandHypo*(h) that generates a list of new hypotheses according to the current hypothesis h and a given confusion network:

```
function ExpandHypo( h )
begin
  for n:=h.pos+1 to N do begin
    foreach w ∈ CFNet(n) do begin
      f := Generate( )
      f.pos := n
      if w = "@" then
        f.words := h.words
        f.score := h.score + λ_C C(n,"@")
        F := ExpandHypo( f )
        Move(G, F)
      else
        f.words := h.words + w
        f.score := h.score +λ_L L(w|h.words)
                   + λ_C C(n, w) + sp*d(h, w) + ip
        Insert (G, f )
      endif
    end
  end
  return G
end
```

where the confidence score $C(w)$ is extended to $C(n,w)$ for using a confusion network, that indicates a logarithmic value of a posterior probability for word w in the n-th confusion set; $d(h,w)$ is a function that returns the number of skipped words between the last word of h and word w.

To improve search efficiency, in *Insert*(H,h) and *Move*(H,F), redundant hypotheses can be removed from the list. If there are multiple hypotheses which have reached the same position and whose last two words are identical, it is enough to retain only one hypothesis which has the maximum score among them in the list. For finding only the best complete hypothesis, it is not necessary to keep such redundant hypotheses. Since a trigram probability applied to the next word of the current hypothesis depends only on the last two words of the hypothesis, only the best hypothesis in the two-word context has a chance to be the best complete hypothesis i.e. the consolidation result in the future.

3. Evaluation experiment

English academic presentation speech in the TED corpus automatically transcribed using the Janus Recognition Toolkit (JRTk) in the IWSpS was used for evaluation experiments. Eight talks were recognized and evaluated by comparing manual consolidations by human.

3.1. ASR system

Eight talks were recognized with an acoustic model trained on 300 hours of Broadcast News (BN) data merged with the close talking channel of meeting corpora. The acoustic model used 42 features and consisted of 300k Gaussians with diagonal covariances organized in 24k distributions over 6k codebooks [4]. The language model (LM) used for the speech recognizer was generated by interpolating a word 3-gram and a class-based 5-gram LM each trained on BN data (160M words) and the proceedings corpus (see Section 2.1.3), and a 3-gram LM based on talks (60k words) by the TED adaptation speakers. The overall OOV rate is 0.3% with a vocabulary size of 25000 words including multi-words and pronunciation variants. The average word error rate of the talks used in this paper is 33.3%.

3.2. Consolidation module

Linguistic score was calculated using BN data (160M words) and the proceedings corpus (17M words). Confidence score obtained from a confusion network by the ASR system was applied. We separated the eight talks into two sets, one is used as a development set and the other is used as a test set, each of which consists of four talks. The best scaling factors for consolidation scores were experimentally determined using the development set. The test set was evaluated based on the best scaling factors.

3.3. Manual consolidation

1-best of ASR output was manually consolidated by deleting disfluent expressions and phrases which have the different meaning from the manual transcription. Table 1 show an example of manual consolidation.

Table 1: An example of a manual consolidation result

REF	which is another topic of interest such as in identifying the sex or the language sex of the speaker the identity of the speaker or the language being spoken
CON	which is another topic of interest such as in identifying the sex [***for*** the language sex of speaker], [***given state*** speaker ***of***]the language being spoken

REF: manual transcription and CON: manually consolidated ASR output, bold and italic words indicate recognition errors and phrases bracketed are removed.

3.4. Evaluation metrics

Automatic consolidation results were compared with manual consolidation results based on word accuracy. To evaluate the performance of deleting misrecognized phrases, ratio of

Table 2: Ratio of extracted words in spoken words and ratio of correctly recognized words in consolidation results.

LectureID	Manual/1-best		Auto/1-best		Auto/ConfNet		ASR
(TestSet)	Ratio%	Prec%	Ratio%	Prec%	Ratio%	Prec%	WACC
dc57s200	64.2	100.0	71.3	88.0	70.6	88.1	67.2
hb64s400	48.1	100.0	62.9	87.7	62.3	87.9	66.6
ro31s400	79.5	100.0	72.7	95.2	72.1	95.2	83.3
yi59s500	57.9	100.0	74.2	86.7	73.4	86.8	65.7
total	66.6	100.0	71.1	90.5	70.4	90.6	72.5

correctly recognized words in automatic consolidation results was calculated. The ratio is equal to a precision.

3.5. Evaluation results

First, we investigate the accuracy of automatic consolidation, and the effectiveness of each score used in the consolidation algorithm. Figure 2 shows the word accuracy of consolidation results derived from 1-best ASR output in the development set and the test set. L, C, and sp indicate the use of linguistic score, confidence score, and skip penalty, respectively. For example, "L+sp" shows the case when only a linguistic score and a skip penalty are used for consolidation. In both sets, "L+C+sp" gave the best accuracy. Hence all scores defined in this paper are effective for consolidation. Although confidence score seems to be dominant compared to the other scores, a high accuracy was not derived using only the confidence score.

Next, we discuss properties of the consolidation results. Table 2 shows the ratio of the number of automatically extracted words in consolidation to the number of spoken words (Ratio%). We also calculated the ratio of the correctly recognized words contained in the consolidation results i.e. a precision (Prec%) to evaluate the performance of removing recognition errors. We compare three cases of manual consolidation from 1-best ASR output (Manual/1-best), automatic consolidation from 1-best ASR output (Auto/1-best), automatic consolidation from confusion networks (Auto/ ConfNet). For reference, word accuracy of 1-best hypothesis in speech recognition (ASR WACC) is also attached in the table.

In Manual/1-best, it is shown that the human subject extracted 66.6% of spoken words in total. Since the subject knew which words were misrecognized, the precision resulted in 100%. On the other hand, as shown in Auto/1-best, the consolidation module selected 71.1% of spoken words, which was a similar value to that of Manual/1-best. In each lecture, however, the ratio was not so similar. Although the human subject tends to extract more words from the ASR output with higher word accuracy, such a tendency did not appear in the result of automatic consolidation.

In this experiment, we could not confirm the efficiency of applying consolidation to confusion networks since the result of Auto/ConfNet is almost the same as that of Auto/1-best. However, in both cases, it is shown that the consolidation method can extract accurately recognized words with high precision above 90%.

4. Conclusions

We proposed consolidation framework for spoken language processing in which reliable phrases maintaining original meanings are attempted to be extracted from ASR output by conducting confidence of ASR and linguistic appropriateness

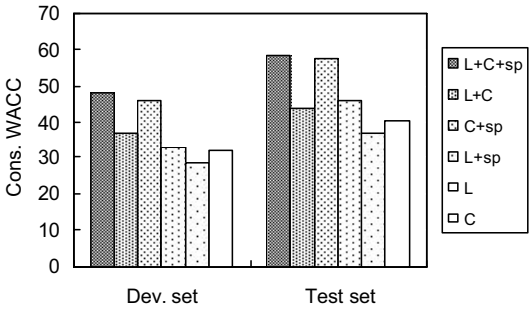

Fig. 2: Word Accuracy of Consolidation Results.

of word concatenation in the extracted phrases. TED speech was recognized and consolidated. Evaluation results show that consolidation results can remove disfluencies and number of recognition errors. Future work involves testing the performance of consolidation for enhancing the total performance of MT and Question Answering.

5. References

[1] K. Zechner, "Summarization of spoken language Chalenges, Methods, and Prospects," Speech Technology Expert eZine, Issue 6 (2002).

[2] C. Hori, S. Furui, R. Malkin, H Yu and A. Waibel, "Automatic Speech Summarization Applied to English Broadcast News Speech," ICASSP (2002).

[3] M. Cettolo, F. Brugnara, and M. Federico, "Advances in the automatic transcription of lectures," ICASSP (2004).

[4] M. Wolfel, S. Burger, "The ISL baseline lecture transcription system for the TED corpus," submitted to Interspeech (2005).

[5] S. Furui, K. Maekawa and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," ASR (2000).

[6] Matthias Honal, Tanja Schultz, "Automatic disfluency removal recognized spontaneous speech –rapid adaptation to speaker-dependent disfluencies," ICASSP (2005).

[7] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," Eurospeech (1997).

[8] L. Mangu, E. Brill and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," Computer, Speech and Language, 14(4):373-400 (2000).