Utterance Verification Incorporating In-domain Confidence and Discourse Coherence Measures

Ian R. Lane[1,2] *and Tatsuya Kawahara*[1,2]

[1]School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[2]ATR Spoken Language Translation Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
ian.lane@atr.jp

Abstract

Conventional confidence measures for assessing the reliability of ASR output are typically derived from "*low-level*" information which is obtained during speech recognition decoding. In contract to these approaches, we propose a novel utterance verification scheme which incorporates confidence measures derived from "*high-level*" knowledge sources. Specifically, we investigate two measures: *in-domain confidence*, the degree of match between the input utterance and the application domain of the back-end system, and *discourse coherence*, the consistency between consecutive utterances in a dialogue session. A joint verification confidence is generated by combining these two measures with an orthodox measure based on GPP (generalized posterior probability). The proposed verification scheme was evaluated on spontaneous dialogue via the ATR speech-to-speech translation system. The two proposed measures were effective in improving verification accuracy.

1. Introduction

Current state-of-the-art speech recognition technology is not robust against acoustic mismatch caused by noise, channel mismatch or speaker variability, or linguistic inconsistencies such as ill-formed or OOD (out-of-domain) input. In order to develop effective spoken language systems based on this technology, it is necessary to detect recognition errors in the ASR output before forwarding it to the back-end natural language processing module. By assessing the confidence of the recognition hypothesis (and individual words within this hypothesis), spoken language systems can generate effective user feedback, applying appropriate recovery strategies depending on the type of error and specific application. For example, a system may confirm only those words with low confidence that are relevant to the current task [1], or may prompt the user to re-speak or re-phrase the entire utterance [2]. To realize such feedback, it is vital to define an effective measure of recognition confidence.

Various approaches have previously been proposed for assessing confidence of ASR output. Feature-based methods, such as [3], assess confidence according to a set of specific features (e.g. word-duration, acoustic and language model back-off, and word graph density). Explicit model-based schemes, such as [4], conduct a likelihood ratio test, comparing the candidate model to a competing model (an anti-model or background model). Posterior-probability-based approaches, including [2] and [5], estimate the posterior probability of a recognized entity (word or utterance) given all competing hypotheses (in an N-best list or word graph). All these approaches, however, basically estimate recognition confidence based on the "*low-level*" information that is available during decoding (for example, normalized acoustic and linguistic likelihoods, and confusability with competing hypotheses). On the other hand, there are apparently knowledge sources outside the ASR framework, such as information about the application domain and knowledge about discourse flow, which have not been well exploited for estimating recognition confidence.

In this paper, we present a novel utterance verification scheme that incorporates such "*higher-level*" knowledge. Specifically, we introduce two confidence measures. The first, *in-domain confidence*, (which was previously proposed in our work on out-of-domain detection [6]) is a measure of match between the input utterance and the application domain of the back-end system. The second, *discourse coherence*, is a measure of the consistency between consecutive utterances in a dialogue session. A joint verification confidence is generated by combining these two measures with a conventional measure based on the GPP (generalized posterior probability) of the ASR output. The effectiveness of the proposed utterance verification scheme was evaluated on spontaneous dialogue via the ATR speech-to-speech translation system [7].

2. Proposed Framework for Utterance Verification

Typical spoken language systems (for example, spoken dialogue, automatic call-routing, and speech-to-speech translation systems) consist of two main sub-systems: an ASR (automatic speech recognition) front-end, which generates a recognition hypothesis for each input utterance, and a NLP (natural language processing) back-end which performs tasks including semantic understanding, dialogue management, and response generation. While conventional approaches to utterance verification [3, 4, 5] typically rely on the information obtained during decoding in the ASR front-end, this paper focuses on incorporating "*high-level*" knowledge sources from the back-end system.

The specific approach proposed in this paper is depicted in Figure 1. The knowledge sources exploited here relate to two very different aspects of spoken language systems and both are expected to be useful for identifying recognition errors that are difficult to detect using only acoustic and linguistic likelihoods. The first measure, *in-domain confidence*, $CM_{\text{in-domain}}(X_i)$, is the degree of match between the in-

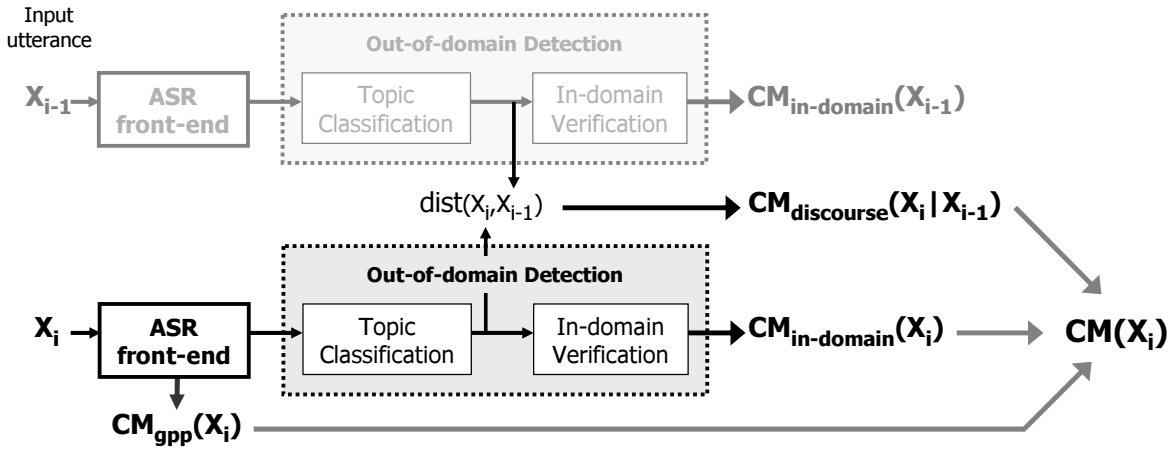

Figure 1: Proposed framework for utterance verification

put utterance against the application domain of the back-end system. This confidence measure will be effective for rejecting recognition errors that are caused by mis-match of domain, and also erroneous hypotheses that do not make sense in terms of the application domain. The second measure, *discourse coherency*, $CM_{\text{discourse}}(X_i|X_{i-1})$, tries to verify the dialogue consistency between consecutive utterances in a dialogue session. This measure will be useful for rejecting erroneous hypotheses that result in inconsistency with the preceding utterance.

A joint confidence measure is realized by combining these two "*high-level*" measures with an orthodox confidence measure based on the GPP (generalized posterior probability [2]) of the recognition output, $CM_{\text{gpp}}(X_i)$. The two proposed confidence measures, *in-domain confidence* and *discourse coherency*, are described in detail in the following sections.

3. In-domain Confidence

The first confidence measure we investigate, *in-domain confidence*, is a measure of match between the input utterance and the application domain of the back-end system. We originally proposed this measure for OOD (out-of-domain) utterance detection (or in-domain verification) in [6]. Its computation is illustrated in Figure 2. We assume that the training set is initially split into multiple topic classes. In this work, topic classes were pre-defined and the training set was hand-labeled appropriately. This data was then used to train the topic classification and in-domain verification models.

In our previous work [6], we observed that most of the OOD utterances detected by this scheme contain speech recognition errors. Moreover, in-domain utterances with significant speech recognition errors tended to be rejected as OOD because the output recognition hypothesis was mis-matched in terms of the application domain. Therefore, the *domain confidence* is also expected to be useful for utterance verification, which rejects recognition errors themselves.

OOD detection is performed in the following steps. First, the recognition hypothesis is transformed to a feature vector (W) consisting of word, word-pair and word-triplet occurrence counts. Next, a topic classification confidence vector $\left(C(t_1|X), \ldots, C(t_m|X)\right)$ is generated by applying support-vector-machine (SVM) based classifiers for each in-domain topic class t_i. An in-domain verification model

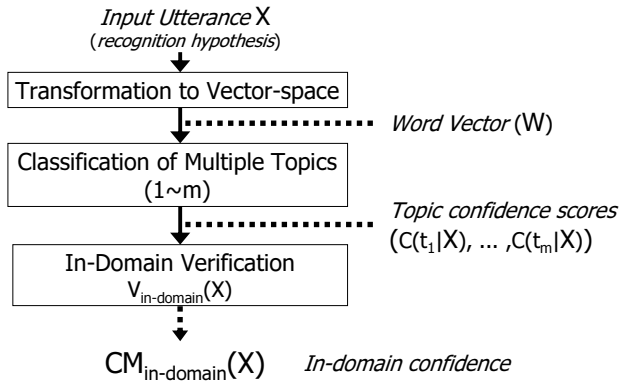

Figure 2: Computation of *in-domain confidence*

$V_{\text{in-domain}}(X)$ (Equation 1) is then applied. The linear discriminant weights $(\lambda_1, \ldots, \lambda_m)$ of this model are trained using only in-domain examples by applying deleted interpolation of topics and the gradient probabilistic decent algorithm, as described in [6].

$$V_{\text{in-domain}}(X) = \sum_{i=1}^{m} \lambda_i C(t_i|W) \quad (1)$$

W: vector representation of input utterance X
m: no. topic classes

Finally, an in-domain confidence score, $CM_{\text{in-domain}}(X)$, is generated by applying a sigmoid function to the resulting verification score.

$$CM_{\text{in-domain}}(X) = sigmoid\Big(V_{\text{in-domain}}(X)\Big) \quad (2)$$

4. Discourse Coherence

The second measure, *discourse coherence*, is based on topic consistency across consecutive utterances. A user's response is typically related to the preceding utterance in the dialogue, either a system prompt in a spoken dialogue system, or the other user's input in a speech-to-speech translation system. If the current utterance is not coherent in terms of dialogue consistency, it is likely that a recognition error occurred in one of these utterances.

To measure *discourse coherence*, we adopt an inter-utterance distance based on the topic consistency between two utterances. It is defined as the Euclidean distance (Equation 3) between the topic confidence vector of the preceding utterance, $\big(C\left(t_1|X_{i-1}\right),\ldots,\left(t_m|X_{i-1}\right)\big)$, and that of the current utterance $\big(C\left(t_1|X_i\right),\ldots,\left(t_m|X_i\right)\big)$.

$$dist\left(X_i,X_{i-1}\right)=\sqrt{\sum_{j=1}^{m}\Big(C\left(t_j|X_i\right)-C\left(t_j|X_{i-1}\right)\Big)^2} \quad (3)$$

A confidence score, $CM_{\text{discourse}}(X_i|X_{i-1})$, is then generated by applying a sigmoid function to this distance.

$$CM_{\text{discourse}}(X_i|X_{i-1})=sigmoid\Big(dist\left(X_i,X_{i-1}\right)\Big) \quad (4)$$

The resulting $CM_{\text{discourse}}\left(X_i|X_{i-1}\right)$ is large when the topic classification results of the two utterances are close, and low when they differ significantly.

5. Joint Confidence by Combining Multiple Measures

The two confidence measures proposed in the previous sections are incorporated into the utterance verification framework by combining them with a standard approach based on GPP (generalized posterior probability [2]), $CM_{\text{gpp}}(X_i)$. A linear weighted model is applied to compute the joint confidence score $CM(X_i)$.

$$\begin{aligned}CM(X_i) \;=&\lambda_{\text{gpp}} * CM_{\text{gpp}}(X_i)\\ &+\lambda_{\text{in-domain}} * CM_{\text{in-domain}}(X_i)\\ &+\lambda_{\text{discourse}} * CM_{\text{discourse}}(X_i|X_{i-1})\end{aligned} \quad (5)$$

where $\lambda_{\text{gpp}}+\lambda_{\text{in-domain}}+\lambda_{\text{discourse}}=1$

A binary verification decision is made by applying a predefined threshold (φ) to the resulting score. The model weights $\big(\lambda_{\text{gpp}},\lambda_{\text{in-domain}},\lambda_{\text{discourse}}\big)$ and decision threshold (φ) are trained to minimize verification errors on a development set.

Table 1: Summary of development and test sets

	Development	Test
# dialogues	270	90
Japanese side		
# utterances	2674	1011
WER	10.5%	10.7%
SER	41.9%	42.3%
English side		
# utterances	3091	1006
WER	17.0%	16.2%
SER	63.5%	55.2%

WER: Word error rate
SER: Sentence error rate

6. Experimental Evaluation

The performance of the proposed utterance verification scheme was evaluated on spontaneous dialogue via the ATR speech-to-speech translation system [7]. This system operates on the travel-conversation domain and performs translation between English and Japanese.

The ATR "basic travel expressions" corpus [8] was used for training of the language models applied during speech recognition and the topic classification and in-domain verification models applied during utterance verification. This corpus consists of 14 topic classes (e.g. accommodation, shopping, transit, etc.) and 400k training sentences for each language side. Development and test sets, which are different from the above corpus, consist of natural spoken dialogue between native English and native Japanese speakers via the ATR speech-to-speech translation system. Dialogue data were collected based on a set of pre-defined scenarios, relating to the travel domain. A summary of these data are shown in Table 1.

6.1. Baseline Speech Recognition Performance

First, the performance of the English and Japanese ASR front-ends were evaluated. ASR was performed using the ATR speech recognition system, ATRASR [9]. Lexicons consisting of 20k and 16k words were applied for the Japanese and English sides, respectively. During recognition word graphs were initially generated by applying a bigram language model. These were then rescored using a trigram language model to obtain the final recognition output. The recognition performance for the Japanese and English dialogue sides are shown in Table 1.

6.2. Evaluation Measures

In speech-to-speech translation tasks, in which there is no definite "*keyword*" set, the most effective method to handle speech recognition errors is to prompt users to re-phrase the entire input (so long as it is in domain). Thus, verification is formulated as rejecting entire utterances if they contain one or more recognition errors[1]. System performance was evaluated based on CER (confidence error rate [4]) (Equation 6). Errors include false acceptance (FA) of incorrectly recognized utterances and false rejection (FR) of correctly recognized utterances.

$$CER=\frac{\#\text{false acceptance}+\#\text{false rejection}}{\#\text{utterances}} \quad (6)$$

[1] We are investigating methods to tolerate trivial errors that do not affect translation quality.

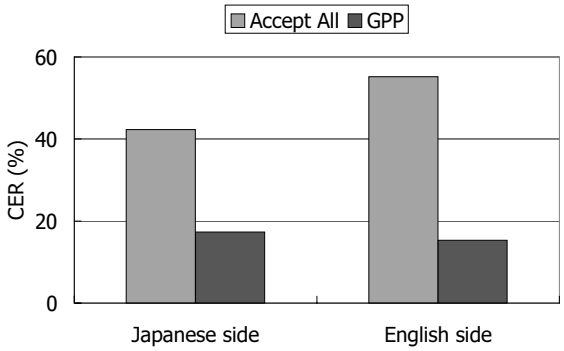

Figure 3: Baseline system performance

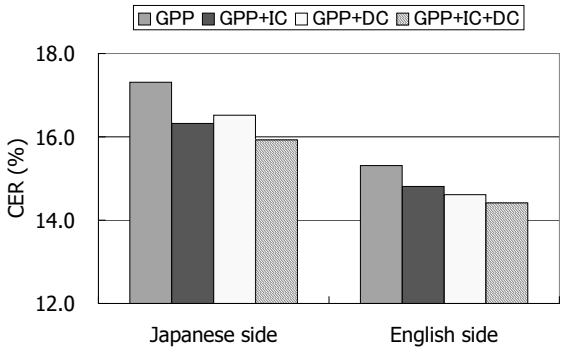

Figure 4: System performance with incorporation of proposed confidence measures

GPP:	Generalized Posterior Probability
IC:	In-domain Confidence
DC:	Discourse Coherence

6.3. Performance of GPP-based Verification

Next, the performance of a GPP-based baseline system was evaluated. In this system, utterance verification was realized by comparing the GPP of the ASR output at the utterance-level to a pre-defined threshold, trained on the development set. The CERs of this system ("*GPP*"), and a reference case where all hypotheses are accepted ("*Accept All*") are shown in Figure 3. The performance of the "*Accept All*" case matches the SER of the respective ASR front-ends. The GPP-based baseline system obtained CERs of 17.3% and 15.3%, respectively for the Japanese and English sides.

6.4. Incorporation of *In-domain Verification* and *Discourse Consistency* Measures

Finally, the proposed utterance verification scheme which incorporates *in-domain confidence* and *discourse coherence* measures was evaluated. The verification performance on the English and Japanese sides for the GPP baseline and when the respective measures were incorporated are shown in Figure 4.

For the Japanese side, incorporating *in-domain confidence* ("*GPP+IC*") and *discourse coherence* ("*GPP+DC*") reduced the CER to 16.3% and 16.5%, respectively, compared to using the GPP measure alone (CER = 17.3%). These correspond to relative reductions in CER of 5.7% and 4.6%, respectively. Incorporating both measures jointly provided a combined reduction in CER of 8.0% (from 17.3% to 15.9%). Similar performance was gained for the English side with a relative reduction in CER of 6.1% (from 15.3% to 14.4%) when both "*high-level*" measures were incorporated.

7. Conclusions

We have investigated a novel utterance verification scheme that incorporates "*high-level*" knowledge into the confidence measure. Specifically two confidence measures were proposed: *in-domain confidence*, the degree of match between the input utterance and the application domain of the back-end system, and *discourse coherence*, the consistency between consecutive utterances in a dialogue session. Evaluation was performed on spontaneous dialogue via the ATR speech-to-speech translation system. The two proposed measures were effective in improving utterance verification accuracy, and the CER was reduced by 8.0% (for the Japanese case) compared to using generalized posterior probability alone.

Acknowledgements: 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".

8. References

[1] T. Misu, K. Komatani, and T. Kawahara. Confirmation strategy for document retrieval systems with spoken dialog interface. In Proc. *ICSLP, pp.45–48*, 2004.

[2] W. K. Lo, and F. K. Soong, Generalized posterior probability for minimum error verification of recognized sentences. In Proc. *ICASSP, pp. 85–89*, 2005.

[3] T. Kemp, and T. Schaff, Estimating confidence using word lattices. In Proc. *EuroSpeech, pp. 827–830*, 1997.

[4] M. G. Rahim, C. H. Lee, and B. H. Juang, Discriminative utterance verification for connected digits recognition. *IEEE Trans. SAP, vol. 5, pp. 266–277*, 1997.

[5] F. Wessel, R. Schluter, K. Macherey, and N. Hermann, Confidence measures for large vocabulary continuous speech recognition. *IEEE Trans. SAP, vol. 9, pp. 288–298*, 2001.

[6] I. Lane, T. Kawahara, T. Matsui and S. Nakamura, Out-of-domain detection based on confidence measures from multiple topic classification. In Proc. *ICASSP, pp. 757–760*, 2004.

[7] T. Takezawa, A. Nishino, K. Takashima, T. Matsui, and G. Kikui. An experimental system for collecting machine-translation aided dialogues. In Proc. *Proc. FIT2003, Vol. 2, pp. 161-162*, 2003.

[8] T. Takezawa, M. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, Towards a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In Proc. *LREC, pp. 147-152*, 2002.

[9] T. Shimizu et al., Spontaneous dialogue speech recognition using cross-word context constrained word graph. In Proc. *ICASSP, pp. 145–148*, 1996.