Evaluating semantic-level confidence scores with multiple hypotheses

B. Thomson, K. Yu, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, S. Young

Engineering Department, Cambridge University, CB2 1PZ, UK

{brmt2, ky219, mg436, sk561, farm2, js532, sjy}@eng.cam.ac.uk

Abstract

In any dialogue manager, confidence scores play a central role in ensuring robust operation. Recently, dialogue managers have attempted to exploit N-best lists of alternatives for the semantics rather than the single most likely interpretation. Each alternative in the N-best list must have an associated confidence score and it is very useful to be able to evaluate the utility of these scored lists independent of the application in which they are used. This paper adapts several traditional metrics for confidence scoring to the context of the N-best semantic hypotheses output by a speech understanding system. An alternative metric, called the Item-level Cross Entropy (ICE), is proposed and is shown to have good theoretical and experimental characteristics. As an example of the use of the metrics, various simple methods for assigning confidences are discussed and evaluated. Of all the metrics tested only the ICE metric provided a consistent monotonic ranking of the various systems.

Index Terms: Robustness, Speech processing, Spoken language understanding, Dialogue systems, Confidence Scoring

1. Introduction

Robustness is a major concern for state-of-the-art dialogue systems. Both speech recognition and semantic errors cause dramatic decreases in performance if they are not dealt with correctly. Any approach to solving this problem requires confidence scores on the semantics received by the dialogue manager. Unless the system has some measure of its belief in what the user is saying, it is unlikely that it will be able to deal with errors effectively.

Current systems typically convert a given speech utterance into exactly one semantic-level output, sometimes called a *dialogue act*. In this situation one can consider confidence scoring as a classification task between correct and incorrect dialogue acts [1]. When the system produces multiple hypotheses for any utterance, a more complex approach may be required.

The use of multiple hypotheses is of particular interest when the dialogue system is based on Partially Observable Markov Decision Processes (POMDP) [2]. These systems can significantly improve robustness by using multiple hypothesised dialogue acts [3]. In a probabilistic framework, such as the POMDP, each confidence scores should estimate the posterior probability of its associated dialogue act. To build a working POMDP-based system, the confidence scores must be evaluated on this basis.

Central to the evaluation of confidence scores is the format used for dialogue acts. This typically depends on the task and there is no generally accepted standard. One approach, which will be used here, is that the act is composed of a series of *semantic items* whose order is unimportant. These semantic items might represent attribute-value pairs or more abstract dialogue act *types*, which distinguish for example whether the utterance

was requesting or giving information. An example utterance, along with reference dialogue acts and semantic items as well as act and item hypothesis lists are given in Table 1. The techniques described in this paper could be extended to other dialogue act formats without significant effort.

Utterance:	I'd like um an expensive hotel please	
Ref. Act:	inform(type=hotel, pricerange=expensive)	
Ref. Items:	(inform, type=hotel, pricerange=expensive)	
Hyp. Acts:	inform(type=hotel, pricerange=expensive)	0.9
	inform(type=hotel, pricerange=inexpensive)	0.1
Hyp. Items:	inform	1.0
	type=hotel	1.0
	pricerange=expensive	0.9
	pricerange=inexpensive	0.1

Table 1: Example utterance with the reference dialogue act (Ref. Act), reference semantic items (Ref. Items), example act hypothesis list (Hyp. Acts) and semantic item hypothesis list (Hyp. Items). Confidences scores are shown in the third column.

This paper adapts various standard metrics for confidence score evaluation to the context of dialogue act confidences. In addition, a new metric based on cross entropy is introduced and shown to have good theoretical and experimental properties. Section 2 introduces the metrics followed in Section 3 by an experimental analysis of their characteristics. An example of using the metrics to distinguish between several simple confidence scoring algorithms is given in Section 4, and Section 5 concludes the paper.

2. Evaluation Metrics

In many ways, evaluation of a semantic parser is the same as the evaluation of any other classifier with multiple outputs. In evaluating a speech recogniser, for example, one compares words with a reference transcription. In the case of a semantic parser either the dialogue acts as a whole or the semantic items are compared.

The use of either exact matches of dialogue acts or partial matches given by counting the matching semantic items give rise to two sets of metrics. Matches at a dialogue act level may be more appropriate if there are strong dependencies between semantic-items whereas item-level matching may give a better overall evaluation of the semantic parser. If the confidences are given only at an act level, they are converted to an item level score by summing the confidences over acts containing the item.

When defining the item-level metrics it is simpler to consider the set of all semantic items rather than just those hypothesised. Semantic decoding then becomes a task of choosing whether the semantic item is correct for a given utterance. In practice, implementation may restrict calculations to the semantic items actually hypothesised or in the reference but concep-

tually matches are compared by summing over all possibilities.

Most of the notation that will be used in definitions is common to all metrics. Starting with an item-based approach, let the number of utterances be U and let W denote the number of all available semantic items. Given $u = 1 \ldots U$ and $w = 1 \ldots W$ let:

$$c_{uw} = \begin{cases} \text{Confidence assigned to the hypothesis that the} \\ w^{th} \text{ semantic item is part of utterance } u, \\ 0 \text{ if none was assigned} \end{cases}$$

$$\delta_{uw} = \begin{cases} 1 & \text{if the } w^{th} \text{ item is in the reference for } u \\ 0 & \text{otherwise} \end{cases}$$

$$N_w = \text{Total number of reference semantic items}$$
$$= \sum_{uw} \delta_{uw}$$

In the example from Table 1, the confidences c_{uw} are all zero except for those corresponding to the semantic items "inform", "type=hotel", "pricerange=expensive" and "pricerange=inexpensive" which are 1.0, 1.0, 0.9 and 0.1, respectively. In the case of metrics defined at an act-level, a slight variation in notation is used. Let the number of hypothesised or reference acts be H and denote for $h = 1 \ldots H$:

$$c_{uh} = \begin{cases} \text{Confidence assigned to the } h^{th} \text{ act being the correct} \\ \text{parse for utterance } u, 0 \text{ if none was assigned} \end{cases}$$

$$\delta_{uh} = \begin{cases} 1 & \text{if the } h^{th} \text{ act is the correct parse for } u \\ 0 & \text{otherwise} \end{cases}$$

2.1. Confidence Weighted Metrics

A simple possibility for evaluating confidence scores is to adjust traditional metrics to take account of the confidence. Where correct items would normally be calculated, one calculates an expected value over the confidence scores. Similarly the number of hypothesised items is replaced with an expected number.

One example of this approach is to convert the semantic error rate into a confidence weighted form. For each act a hypothesised for utterance u, the items contained in a are matched with the items contained in the reference and the sum of the item substitutions, deletions and insertions are calculated and denoted e_{ua}. The confidence weighted semantic error rate is then:

$$\text{WSER} = \frac{1}{N_w} \sum_{u,h} c_{uh} e_{uh} \qquad (1)$$

When using confidence-weighted metrics for evaluation, it soon becomes obvious that good confidence scores are not necessarily reflected in an improved score. As shown in section 3, confidence weighted error rates actually increase with the number of hypotheses. This is counter-intuitive since the larger list has more information and should perform better.

A theoretical explanation for this issue comes by examining the choices made by the confidence scorer. Suppose that the scorer has some beliefs $\mathcal{B}$ about the semantics of each utterance and aims to optimise the expected value of the metric under its beliefs. Under the error rate metric this corresponds to optimising:

$$\mathbb{E}\left(\sum_{u,h} c_{uh} e_{uh} \mid \mathcal{B}\right) = \sum_{u,h} c_{uh} \mathbb{E}(e_{uh} \mid \mathcal{B}) \qquad (2)$$

Given the constraints $\sum_h c_{uh} = 1$ and $c_{uh} \geq 0$, the optimum is achieved by setting $c_{uh} = 1$ for the hypothesis with minimum

expected error. Added hypotheses will always result in worse expected semantic error rates. This suggests severe deficiencies in the metric as no credit is being given to the accuracy of the confidence scores. Confidence weighted recall, precision or F-scores can also be defined but suffer from similar problems.

2.2. NCE scores, Oracle rates and other metrics

One common metric for evaluating speech recognition confidences is the normalized cross entropy (NCE). This was the method used for several NIST evaluations and details of its application to other natural language processing tasks may be found in [4]. An equation for the item-level form of NCE is:

$$\text{NCE} = \frac{H_{\text{base}} + \sum_{u,w} \log(\delta_{uw} c_{uw} + (1 - \delta_{uw})(1 - c_{uw}))}{H_{\text{base}}} \qquad (3)$$

where $H_{\text{base}} = n_c \log p_c + (N_h - n_c) \log(1 - p_c)$, $p_c = \frac{n_c}{N_h}$, n_c is the number of correct semantic items from this list of hypotheses and N_h is the number of hypothesised items (hypothesised items are those with $c_{uw} > 0$).

The reason for normalising by H_{base} is to adjust for the overall probability of correctness to enable comparisons between data sets. H_{base} gives the entropy that would be obtained by simply using the constant probability, p_c. This normalisation term, however, depends on the number of hypothesised items. The score can be increased by simply adding more hypothesised items with very low probability. NCE is thus a suitable metric for evaluating the accuracy of probability estimates given a set of hypotheses, but it does not necessarily test the overall correctness of the output.

A useful measure of correctness is the oracle error rate, which measures the error rate that would be achieved if an oracle chose the best option from each hypothesised list of dialogue acts. This gives an upper bound on the error that could be achieved for a given list of hypotheses. Unfortunately, it is clearly not appropriate as an overall metric since confidence scores are ignored.

Another commonly used tool for the evaluation of confidence scores is the receiver operating characteristic (ROC) curve [4]. One considers a classifier based on the confidence score which accepts or rejects hypotheses depending on a confidence threshold. The ROC curve then plots the number of correct rejections and acceptances. The problem with this is that only the first hypothesis and its confidence is ever evaluated.

2.3. Cross Entropy

The traditional metrics discussed above give a way to evaluate either the confidence scores or the overall correctness, but not both. An ideal metric should incorporate both factors, as well as giving a good indication of the effect on dialogue performance. This leads to the proposal of a new metric, based on the cross entropy between the probability density from the confidences and the optimal density given by delta functions at the correct values. This is very similar to the NCE metric, but does not normalise for the average probability of correctness. An Item-level Cross Entropy (ICE) is defined below, although a similar metric could be defined for act-level evaluations.

$$\text{ICE} = \frac{1}{N_w} \sum_{u,w} -\log(\delta_{uw} c_{uw} + (1 - \delta_{uw})(1 - c_{uw})) \qquad (4)$$

Assuming that the total number of reference items N_w is fixed, consider the decisions that the confidence scorer makes

as was done in section 2.1. The scorer must aim to optimise the expected value of the metric:

$$\mathbb{E}(\text{ICE}|\mathcal{B}) = \frac{-1}{N_w} \sum_{u,w} [p_{uw} \log(c_{uw}) + (1-p_{uw}) \log(1-c_{uw})] \tag{5}$$

Differentiating with respect to c_{uw} gives

$$\frac{p_{uw} - c_{uw}}{c_{uw}(1 - c_{uw})} = 0 \tag{6}$$

and so the minimum is achieved when $c_{uw} = p_{uw}$. When substituting this optimum into 5, the expected value of the metric is the average entropy of the beliefs $\mathcal{B}$. The metric therefore penalises systems for bad confidence scores as well as giving credit for bolder predictions.

3. Experimental analysis of metrics

An experimental evaluation of the metrics discussed above was completed on a corpus of 648 dialogues recorded during a user trial of various dialogue managers [5]. Users were asked to imagine themselves in an unknown town and interacted with the dialogue managers to find a hotel, bar or restaurant that matched some predetermined constraints. The corpus contains around 5800 utterances with semantic annotations. During experiments three levels of noise were added to evaluate the effects of noise although the original clean signal was recorded. The noise signal was directly added online to the raw waveform before speech recognition and resulted in overall signal to noise ratios of 35.3db (zero noise), 10.2db (medium noise) and 3.5db (high noise).

Confidence scoring is implemented by first constructing the confusion network from lattices output by the speech recogniser[6]. Each word arc in the confusion network has a log posterior associated which is used in a dynamic programming search to construct an N-Best list. The summation of these log posteriors is called the *inference evidence* and after exponentiating and renormalising is used for the sentence-level score. Confidences on the semantics are calculated by summing the sentence-level scores for all sentences which are parsed as the same dialogue act.

Table 2 shows a comparison of the various metrics on offline transcription experiments. In offline experiments, the confidence scorer described above was compared against an alternative approach which simply parses each sentence, chooses the dialogue act with the most compatible sentences and assigns it probability 1 (Const). The last two lines in the table show how the confidence weighted semantic error rate actually improves with a smaller number of speech to text (STT) hypotheses, contrary to the intuitive trend of higher performance with more information. Comparing the second and last lines shows how the oracle error rate is also inadequate as a metric since the scores are equal even though the confidence scores are completely different. Using the ICE metric gives a constantly decreasing error, as would be expected.

The problems involved with using the NCE score are more difficult to observe, but can also be seen in the table. Adding noise with the original confidence scorer decreases performance more than using a less effective confidence scorer. With the NCE, the effect of the inaccurate confidence scores far outweigh the lower oracle error rate. Similarly, decreasing the number of speech hypotheses is deemed worse than adding noise. This is despite the fact that the noise has significantly increased the oracle error rate.

Noise (db)	Conf. Cal.	N-Best	Metric			
			SER	ORA	NCE	ICE
10.2	InfEv	100	—	16.4	0.352	1.737
35.3	Const1	100	—	7.9	-0.641	1.706
35.3	InfEv	5	19.2	12.2	0.212	1.103
35.3	InfEv	100	20.6	7.9	—	0.941

Table 2: Comparison of evaluation metrics for speech understanding with multiple hypotheses. The table compares the Confidence Weighted Semantic Error Rate (WSER), Oracle Error Rate (ORA), Normalised Cross Entropy (NCE), and Item-level Cross Entropy (ICE) metrics. Most experiments use the information evidence approach (InfEv) which is compared against a less effective scorer (Const). All metrics except NCE show improvements as decreases in the metric value.
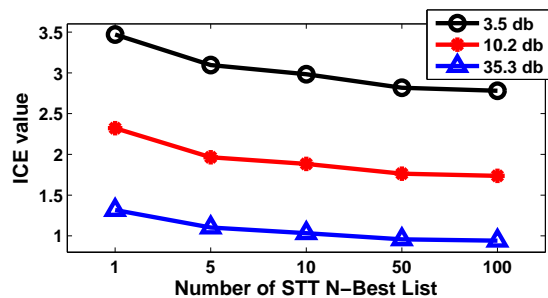

Figure 1: ICE scores obtained by varying the amount of noise and the number of STT hypotheses.

The ICE metric has some further desirable properties which are worth mentioning. Firstly, the ICE metric degrades with added noise and improves when the number of STT hypotheses increases. Experimental results on the corpus are shown in Figure 1

The second desirable feature of the metric is that it gives a good indication of the effect of spoken language understanding on dialogue performance. Since the corpus consists of data recorded from interactions with various dialogue managers one can plot dialogue performance as a function of the proposed metric for the different dialogue managers. Performance is measured by giving 20 points for a successful dialogue, 0 for an unsuccessful one and subtracting 1 point for each dialogue turn. The ICE metric is continuous so it is easier to visualise by using a regression model. For this purpose a Gaussian Process regression was used because of its flexibility [7][1].

Two of the systems in the trial are based on the Bayesian Update of Dialogue State model, which is derived from the Partially Observable Markov Decision Process model (POMDP). Two other systems are based on the more traditional finite state Markov Decision Process (MDP). On both simulations and in this experiment the POMDP models have been shown to outperform the MDP model [5]. As shown in Figure 2 this separation of dialogue performance is maintained when using the ICE metric on the observed data with real users. In simulation experiments from [5] the simulated error channel made use of confusion rate to vary the amount of error. Although not shown here, other experiments suggest that the ICE is an appropriate substitute for real experiments since simulation values are strongly correlated.

---

[1] For GP regression we used a Matern class covariance function with $\nu = \frac{5}{2}$ and type II maximum likelihood estimation of parameters
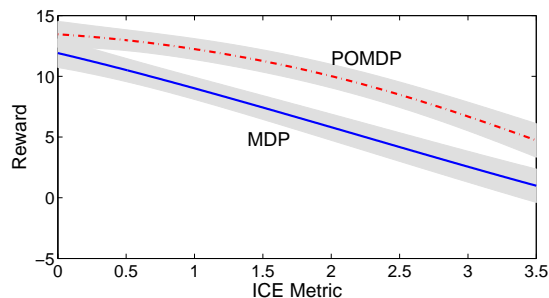
Figure 2: Mean reward for real-users as a function of the ICE metric estimated using a Gaussian Process regression model. The shaded region shows one standard error on each side of the mean.
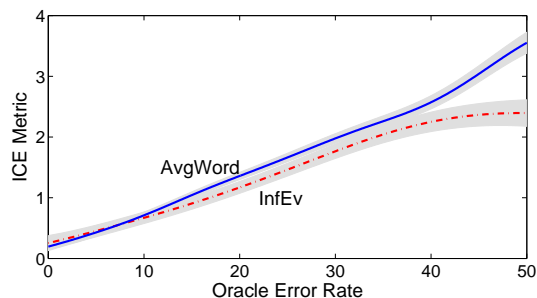

Figure 3: ICE metric for different confidence scoring algorithms as a function of oracle error rate as estimated by the Gaussian Process regression model. The shaded region shows one standard error on each side of the mean.

4. Comparison of confidence scorers

An example comparison using the metrics above was completed using several simple confidence scoring techniques. Each STT hypothesis is assigned a sentence-level confidence score which is normalised so that the sum over the N-best list is 1, and the sentence is then passed to the semantic parser. The parser determines the most likely dialog act for each sentence, then groups together sentences which produce the same dialog act and adds confidence scores are added. Three methods were evaluated for the STT sentence-level confidence score:

Const Constant value for each STT hypothesis

AvgWord Average of all word-level confidence scores

InfEv Exponentiated inference evidence

Table 3 shows a comparison of the confidence scoring algorithms on the corpus from the previous section. All results use a 5-best list of speech recognition hypotheses. On all metrics, the inference evidence approach achieves the highest performance. The ICE and NCE metrics show that using multiple hypotheses is better than using only one.

Conf. Cal.	Metric		
	WSER	NCE	ICE
Const.	22.3	0.143	1.453
AvgWord	22.3	0.160	1.138
InfEv	19.2	0.212	1.103
InfEvTP1	16.6	-0.523	1.331

Table 3: Comparison of different confidence scoring methods. The last line (InfEvTP1) uses the inference evidence for STT confidences but after grouping together equivalent sentences keeps only the most likely dialog act and assigns it a confidence of 1.

While the offline experiments from Table 3 are useful it is important to check results with online experiments as well. This was done by comparing the results from the above corpus, which used inference evidence, against a second user trial, which used the average word score as the STT confidence. This trial contained 504 dialogues and around 4000 utterances.

Figure 3 shows a comparison of the different ICE scores as a function of oracle error rate. The figure again uses a Gaussian Process regression model to obtain estimates of the mean value for each confidence scorer. The figure indicates that the inference evidence approach is more effective than average word score, although more data is required for conclusive results.

5. Summary

This paper has shown how various traditional metrics for evaluating confidence scores can be adapted for evaluating the confidence scores for N-best dialogue act recognition. A new metric, called the Item-level Cross Entropy (ICE), was proposed and shown to give a consistent performance ranking for both the confidence scores and the overall correctness of the system. Using the metrics, various confidence scoring algorithms were evaluated.

Future work will focus on developing better techniques for calculating confidence scores. In the meantime, the ICE metric provides a very useful tool for evaluating our speech understanding performance and for comparing confidence scores.

6. Acknowledgements

7. References

[1] P. Carpenter, C. Jin, D. Wilson, R. Zhang, D. Bohus, and A. Rudnicky, "Is this conversation on track?," in *Eurospeech-2001*, Aalborg, Denmark, 2001.

[2] J. Williams and S. Young, "Partially Observable Markov Decision Processes for Spoken Dialog Systems," *Computer Speech and Language*, vol. 21, no. 2, pp. 231–422, 2007.

[3] B. Thomson, J. Schatzmann, and S. Young, "Bayesian update of dialogue state for robust dialogue systems," in *ICASSP*, Las Vegas, NV, 2008.

[4] S. Gandrabur, G. Foster, and G. Lapalme, "Confidence estimation for NLP applications," *Transactions on Speech and Language Processing*, vol. 3, no. 3, pp. 1–29, oct 2006.

[5] B. Thomson, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, K. Yu, and S. Young, "User study of the bayesian update of dialogue state approach to dialogue management (submitted)," in *INTERSPEECH*, 2008.

[6] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *ICASSP*, 2000.

[7] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning, Ch 2*, MIT Press, 2006.