

Accurate Probability Estimation of Hypothesised User Acts for POMDP Approaches to Dialogue Management

Paul A. Crook

HCRC/ICCS, School of Informatics
University of Edinburgh, UK
pcrook@inf.ed.ac.uk

Oliver Lemon

HCRC/ICCS, School of Informatics
University of Edinburgh, UK
olemon@gmail.com

Abstract

Current Partially Observable Markov Decision Process (POMDP) Reinforcement Learning (RL) approaches to dialogue management require accurate estimates of the probability of each hypothesised user semantic act given the input user’s utterance. In previous work, the probabilities for each hypothesised user act have been approximated using the Automatic Speech Recogniser’s (ASR’s) confidence score for each item in an N-best list. In this paper we examine this approximation and propose a better one. Our results suggest that this approximation does not in general hold. The validity of the assumption also depends on the choice of “confidence score” measure. Our results provide independent support for a previous result which showed that confusion network-based inference evidence (IE) scoring should be preferred for POMDP dialogue management. We also propose a simple remapping of IE scores which should improved probability estimates and show how our analysis can be used to build data driven Automatic Speech Recogniser - Spoken Language Understanding (ASR-SLU) models for simulated user training.

1 Introduction

Accurate confidence scoring from automatic speech recognisers (ASR) is clearly an important ingredient of many spoken dialogue systems. In particular, current approaches in applying Partially Observable Markov Decision Process (POMDP) Reinforcement Learning (RL) to dialogue management (Henderson and Lemon, 2008; Thomson et al., 2007; Thomson

et al., 2008; Gašić et al., 2008) use ASR confidence scores as an estimation of the probability of a user’s hypothesised input (hypothesised user semantic act) given the user’s observed speech; see section 2 for details.

Testing this estimation is important as errors in the probability estimate that it provides will have knock on effects in the updating of the belief state (which is maintained over all possible states of the conversation). Thus effecting the optimality of policies learnt by POMDP RL dialogue systems.

In this paper we examine this approximation for the HTK speech recogniser using a transcribed corpus from the Town Info domain (Lemon et al., 2006).

Two measures of confidence scoring are considered. A standard average word confidence score for each phrase and confusion network based Inference Evidence (IE) scores (Gašić et al., 2008; Mangu et al., 2000).

The remainder of this paper is set out as follows. Section 2 describes relevant literature in the domain of POMDP RL for dialogue systems and positions our work with respect to other approaches to ASR hypothesis re-ranking/classification. Section 3 set out the hypothesis we aim to test and our methodology. Section 4 presents the results of our analysis and section 5 sets out our interpretation of these results and how they support or falsify the hypothesis under test. Section 5 sets out a proposal for improved probability estimation and also indicates how the data collected in this work can be used to set up a data driven simulated Automatic Speech Recogniser - Spoken Language Understand-

ing (ASR-SLU) model for training of POMDP RL dialogue managers. Finally section 6 summaries our results and sets out a plan to validate the proposed improvement to probability estimation for user’s semantic acts given utterances.

2 Background

2.1 POMDP RL approaches to dialogue management

Henderson and Lemon (2008) presents a POMDP formulation for dialogue management where a mixture of MDP states is maintained which represents the full POMDP space. The belief state update for this Mixture Model POMDP approach is

$$b_t \approx \sum_{h_t^j} \sum_{r_{t-1}^i} \frac{1}{P(h_t^j)Z(V_t)} p_{t-1}^i P(h_t^j | a_{t-1}, r_{t-1}^i) P(h_t^j | u_t) b(f(r_{t-1}^i, a_{t-1}, h_t^j)) \quad (1)$$

where b_t is the updated belief state, $P(h_t^j)$ is a prior over ASR-SLU outputs, $Z(V_t)$ is a normalisation constant, p_{t-1}^i is the probability associated with the MDP state r_{t-1}^i at time $t - 1$, h_t^j is the hypothesised j^{th} user act at time t , $P(h_t^j | a_{t-1}, r_{t-1}^i)$ is the *user model* which predicts the likelihood of h_t^j give the previous system action a_t and the previous MDP state r_{t-1}^i , $P(h_t^j | u_t)$ is the probability of the hypothesised user act h_t^j given the user utterance u_t and $b(f(r_{t-1}^i, a_{t-1}, h_t^j))$ is the belief state generated by applying a dialogue update $f(\cdot)$ to a previous MDP state r_{t-1}^i , system action a_{t-1} and hypothesised user act h_t^j .

Understanding the full details of this belief update equation is not important to this paper. Interested readers can find more details in Henderson and Lemon (2008). What is important is to notice the term $P(h_t^j | u_t)$ in equation 1. This term, which is the *probability of the hypothesised user act h_t^j given the user utterance u_t* , is taken as the ASR-SLU confidence score in Henderson and Lemon (2008).

Similarly in the HIS system (Gašić et al., 2008; Young et al., 2007), which uses a partition splitting to represent POMDP states, the belief update equation is:

$$b'(p', a'_u, s'_d) = k P(o' | a'_u) P(a'_u | p', a_m) \sum_{s_d} P(s'_d | p', a'_u, s_d, a_m) P(p' | p) b(p, s_d) \quad (2)$$

where $b'(p, a'_u, s'_d)$ is the updated belief space partition, k is a normalisation constant, $P(o' | a'_u)$ is the *observation model*, i.e. probability of observing o' given user action a'_u , $P(a'_u | p', a_m)$ is the *user action model*, i.e. likelihood of user action a'_u given partition p' of the state space and previous system action a_m , $P(s'_d | p', a'_u, s_d, a_m)$ is the *dialogue model*, i.e. probability of dialogue state s'_d given partition p' , user act a'_u , previous dialogue state s_d and system action a_m , $P(p' | p)$ is the probability of partition splitting and this term together with the belief term $b(p, s_d)$ form a belief refinement step.

Again full details about this update equation are not important. The important factor as far as this paper is concerned is the *observation model*, $P(o' | a'_u)$, which in Gašić et al. (2008) is approximated by the “normalised distribution of confidence measures output by the speech recognition system”.

It is this approximation, $P(h_t^j | u_t) \approx c_t^j$ for the Mixture Model POMDP and $P(o' | a'_u) \approx c_t^j$ for HIS (where c_t^j is the ASR-SLU confidence in the j^{th} hypothesised semantic user act at time t) that we test in section 3.

2.2 Previous work on re-ranking and classifying ASR hypotheses

Several groups have worked on classifying or re-ranking ASR hypotheses. However, none of them actually compute the *probability* that a hypotheses is correct, which is what the latest POMDP models require.

Litman et al. (2000) use acoustic-prosodic information extracted from speech waveforms, together with information derived from their speech recogniser, to automatically predict misrecognised turns in a corpus of train-timetable information dialogues. Walker et al. (2000) use a combination of features from the speech recogniser, natural language understanding, and dialogue manager/discourse history to classify hypotheses as correct, partially correct, or misrecognised. However, both Litman et al. (2000) and Walker et al. (2000) consider only single-best recognition results and thus use their classifiers as “filters” to decide whether the best recognition hypothesis for a user utterance is correct or not. Our work goes further in that we compute probabilities of correctness for hypotheses.

Gabsdil and Lemon (2004) similarly perform re-ordering of n-best lists by combining acoustic and pragmatic features. Their study shows that dialogue features such as the previous system question and whether a hypothesis is the correct answer to a particular question contributed more to classification accuracy than the other attributes. In POMDP approaches such attributes will be computed by the user model as part of the belief update process.

Jonson (2006) classifies recognition hypotheses with labels denoting acceptance, clarification, confirmation and rejection. These labels were learnt in a similar way to Gabsdil and Lemon (2004) and correspond to varying levels of confidence, being essentially potential directives to the dialogue manager. Apart from standard features Jonson includes attributes that account for the whole N-best list, i.e. standard deviation of confidence scores. Another recent approach by Lemon and Konstas (2009) shows that user simulation predictions can be used to improve this type of hypothesis classification. Again, none of these approaches computes the probabilities required by POMDP models.

3 Experiment

The *hypothesis* that this experiment sets out to test is:

that the probability of a hypothesised user semantic act given the user’s utterance, $P(h_t^j|u_t)$ or $P(o^j|a_u^j)$, can be approximated as the confidence score returned by the ASR for each phrase in an N-best list, e.g. $P(h_t^j|u_t) = c_t^j$ where c_t^j is the confidence score of j^{th} hypothesis at time t .

If this hypothesis holds we would expect that for a plot of the probability of correctly parsed user semantic acts versus confidence score, the data points should lie close to line with intercept 0.0 and gradient 1.0. Further, we would expect a test for linear correlation to show that the probability of correctly parsed semantic acts is highly correlated with confidence score values.

Given a transcribed corpus of users interacting with a dialogue system we compared the output of an ASR-SLU against parsed transcriptions. The corpus consisted of 2,076 transcribed utterances. For

each user’s utterance we parsed the transcription and compared it against parsed N-best hypotheses. The N-best hypotheses were output by the ASR when provided with the original audio. A count was made of matching and non-matching hypothesis over the full range of confidence scores. The count of matches versus non-matches gives a approximate measure of the correctness of each hypothesis output by the ASR-SLU combination. We then examined how the probability of ”correctness”, (*i.e.* the probability of a matching parsed user semantic act) varied against confidence score.

3.1 Methodology

We used the HTK speech recogniser, transcribed a corpus from Lemon et al. (2006) and two parsers; the grammar-based GF Parser (Ranta, 2004) and the statistical “Beast” Parser (Meza-Ruiz et al., 2008). The audio files of user utterances were fed into ATK which then generated a N-best list of hypothesised phrases (where N=3) for each user utterance. For each hypothesised phrase two confidence measures were computed by ATK. An average word confidence score for each phrase and a confusion network based IE score (Gašić et al., 2008; Mangu et al., 2000). We examined both of these measures to see if either fulfils the assumption under test.

Both the output phrases from ATK *and* the human transcription were parsed using GF and Beast Parsers. The GF parser is used preferentially with the system falling back to the Beast parser should GF be unable to provide a parse. The output parsed semantics are of the form `[[dialogue_act_type] [value]]`, e.g. `[[ask_option_yes] [yes]]`, `[[spotting_hotel] [hotel]]` or `[[ask_location] [centre]]`. Note that a single phrase can contain two or more pieces of information, e.g. `“[[ask_option_yes, spotting_hotel] [yes, hotel]]”`. Simple post processing of each parse was carried out, removing “null” semantic pairs (*i.e.* `[[] []]`), standardising equivalent terms, *i.e.* “dont care” and “doesnt matter”, and flattening nested brackets. This was done to avoid unnecessary misclassification of parses that would otherwise match.

The results were binned using the confidence

score assigned to each hypothesis. 101 bins were used, each bin had a width of 0.01 with the first bin centred on 0.0 and the last bin centred on 1.0. The bins were also divided based on hypothesis number and confidence score measure used, thus allowing the effects of hypotheses number and confidence score metric to be reported. For each bin a count of matching ($S_{match}^{b,j,m}$) and non-matching ($S_{nonmatch}^{b,j,m}$) was recorded. Where $b \in [0, 0.01, \dots, 1.0]$ is the confidence score bin, $j \in [1, 2, 3]$ is the hypothesis number, $m \in [\text{avg.word, i.e.score}]$ is the confidence score metric used.

From these counts we can then compute the variation in probability of a match between the parsed transcribed utterance and the parsed ASR output as the confidence score varies. The probability of a match is computed as:

$$P_{match}^{b,j,m} = \frac{S_{match}^{b,j,m}}{S_{match}^{b,j,m} + S_{nonmatch}^{b,j,m}} \quad (3)$$

Bins containing less than ten samples, *i.e.* where $S_{match}^{b,j,m} + S_{nonmatch}^{b,j,m} < 10$, were excluded.

4 Results

Figure 1 shows plots of the probability of parsed ASR semantic acts matching the parsed transcribed semantic acts as confidence score is varied. Plots are shown for each of the 3 hypotheses in the N-best list and also for the two confidence score measures. The left hand column presents the IE confidence measure and the right hand column presents the average word confidence measure.

Table 1 presents correlation results for the same data as plotted in figure 1. Pearson’s linear correlation was computed to indicate the correlation between confidence score and the probability of matching parsed user semantic acts.

5 Discussion

If the hypothesis that we are testing (section 3) held we would expect the points plotted in figure 1 to lie along a line with intercept 0.0 and gradient 1.0. Such a line is plotted as a dashed line across each figure. As is apparent from figure 1 few of the plots seem to exhibit this relationship. The closest is figure 1(a) which is the 1st hypothesis in the N-best list and uses IE as the confidence score metric. This

hypothesis number	IE	av. word conf.
1st	0.8439	0.6550
2nd	0.4185	-0.2114
3rd	0.3202	0.2090

Table 1: Pearson’s linear correlation coefficient for the data points plotted in figure 1. Correlation is computed between the confidence score and the probability of matching hypotheses. A correlation coefficient of 1.0/-1.0 indicates perfect correlation/inverse correlation, a value of approaching 0.0 indicates little correlation.

observation is supported by the correlation results shown in table 1. For Pearson’s linear correlation a coefficient of 1.0 or -1.0 indicates perfect correlation/inverse correlation, a value of approaching 0.0 indicates little correlation. The best correlation of 0.8439 is again found for the first hypothesis when using IE scoring. The correlation for the second and third hypotheses, especially for average word confidence, is very low.

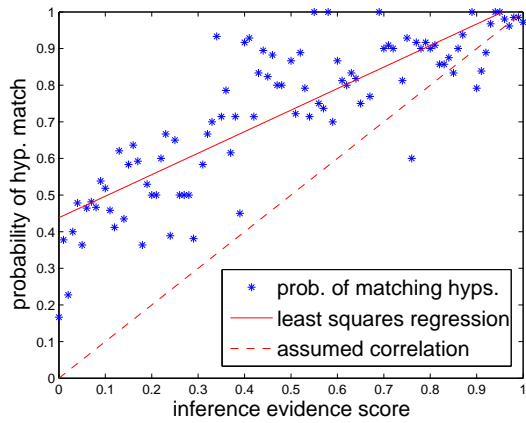
5.1 Experimental hypothesis

The results shown in figure 1 and table 1 indicate that the assumption commonly made in POMDP RL approaches to dialogue management that “*the probability of a hypothesised user semantic act given the user’s utterance, $P(h_t^j|u_t)$ or $P(o'|a_u')$, can approximated as the confidence scores returned by the ASR for each phrase in the N-best list*” is in general not a sound assumption. Based on this data it could possibly be seen as reasonable for the first hypothesis in an N-best list, especially when IE is used as the confidence measure. For the subsequent hypotheses in the N-best list this approximation clearly does not hold. On this corpus IE based metric underestimates the probability of the user’s semantic act given the user’s utterance and average word confidence metric over-estimates the probability of the user’s semantic act given the user’s utterance.

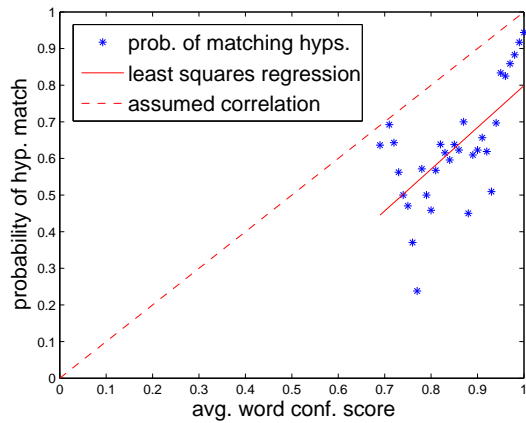
However these results do suggest an approach for improved approximation for this domain.

5.2 Improved probability estimation

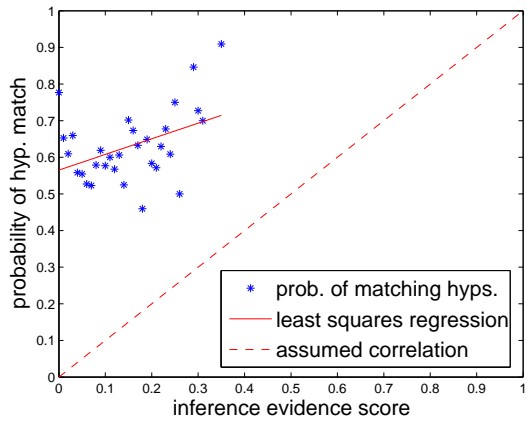
Linear regressions were computed on the data and plotted in figure 1. The resulting best fit lines (minimising the mean squared error) are plotted in fig-



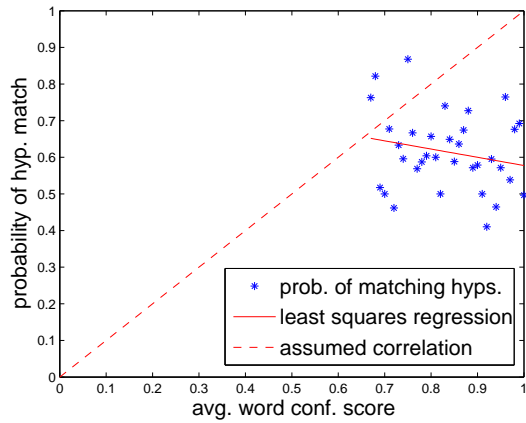
(a) 1st hypothesis, IE score



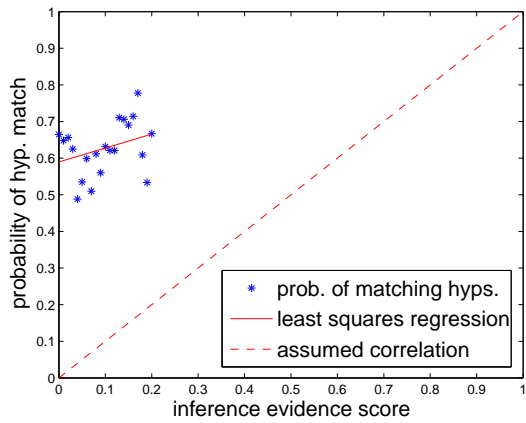
(b) 1st hypothesis, av. word conf.



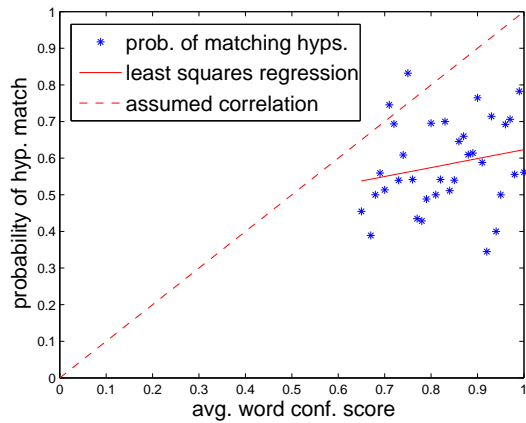
(c) 2nd hypothesis, IE score



(d) 2nd hypothesis, av. word conf.



(e) 3rd hypothesis, IE score



(f) 3rd hypothesis, av. word conf.

Figure 1: Probability of matching parsed hypotheses plotted against confidence score. Left-hand column is using IE as the confidence score measure. Right-hand column is using average word confidence as the confidence score measure. Data was binned according to its confidence score in 101 bins. Starting bin centred on 0.0 and last bin centred on 1.0, each bin is 0.01 wide. Bins where the number of samples was less than 10 are excluded. A linear least squares regression was performed on the plotted points.

ure 1 as solid red lines. Table 2 lists the gradient, intercept and mean squared error for each line.

For the domain in which this corpus was collected, these linear regressions could be used to re-map confidence scores onto a more accurate estimation of the probability of the user’s semantic act. As part of future work we will implement such an approach and test it in a full POMDP system.

Overall correlation results in table 1 suggests that the IE scoring should be preferred when performing a linear remapping of confidence scores to probabilities. For each hypothesis number in the N-best list there is a higher correlation for IE scores than there is for the equivalent hypothesis number using average word confidence scores. This finding is in line with a previous study looking at meta-metrics for determining good ranking metrics of N-best lists, (Thomson et al., 2008), which also suggested that IE scoring should be preferred for POMDP dialogue system approaches.

5.3 Simulated ASR-SLU modelling

Because of the large number of example dialogues required for training simulated dialogues are often generated to train POMDP RL dialogue systems. The above results can also be used to build data driven ASR-SLU model for POMDP training. For our preferred metric of IE scores ($m = i.e. score$), table 3 and figure 2 summarise the raw counts $S_{match}^{b,j,m}$ and $S_{nonmatch}^{b,j,m}$ and show how the distribution of matching and non-matching hypotheses varies with confidence score.

Table 3 shows the likelihood of each parsed ASR hypothesis matching the parsed transcription on this corpus. It shows that in the N-best list the 1st hypothesis is typically correct 74% of the time and the 2nd and 3rd hypotheses are only correct around 62 – 63% of the time. Figure 2 then shows how the distribution of matching and non-matching parsed hypotheses varies with confidence score. These distributions have been normalised such that the area under the matching and non-matching curves are 1.0.

Table 3 can be used in conjunction with the distributions shown to create a data driven simulated ASR-SLU model which should provide better training for POMDP RL dialogue systems.

Such a simulation would work as follows. A sim-

ulated user generates an act in based on its goal and previous system actions. This act would be correctly or incorrectly parsed by the simulated ASR-SLU in accordance with the frequencies set out in table 1. For example, the first hypothesis in a simulated N-best list contains the correct user semantic act 74% of the time. If a “correct” first hypotheses has been generated its simulated *IE confidence score* would be drawn from the “matching distribution” (shown as blue solid line) in figure 2(a). Similarly, for an “incorrect” parse, which occurs 26% of the time, the simulated IE score would be drawn from the “non-matching distribution” (shown as a red dashed line) in figure 2(a).

6 Conclusions & Future Work

We have examined the assumption that for a POMDP RL system the probabilities for each hypothesised semantic user act given the user’s utterance can be approximated as the confidence scores returned by the ASR for each phrase in the *N*-best list.

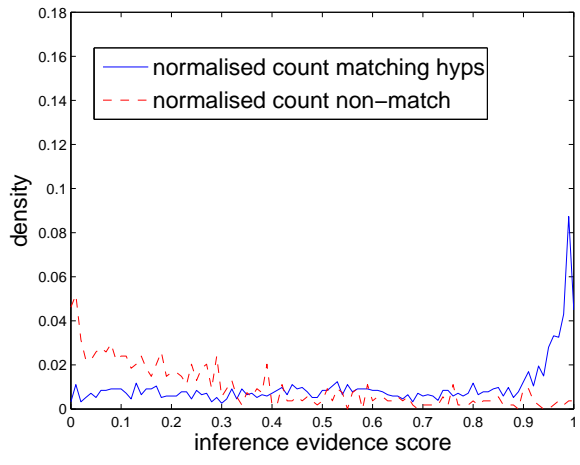
Our results suggest that in general this approximation does not hold, though it could be partly justified for the first item in an N-best list, especially when the confidence score metric used is IE score as computed over a confusion network (Gašić et al., 2008; Mangu et al., 2000).

We have suggested that a simple linear remapping based on a regression analysis may provide improved probability estimation. This approach is as yet untested. We plan to implement this remapping using a POMDP RL system and test if this probability estimation improves performance with real users.

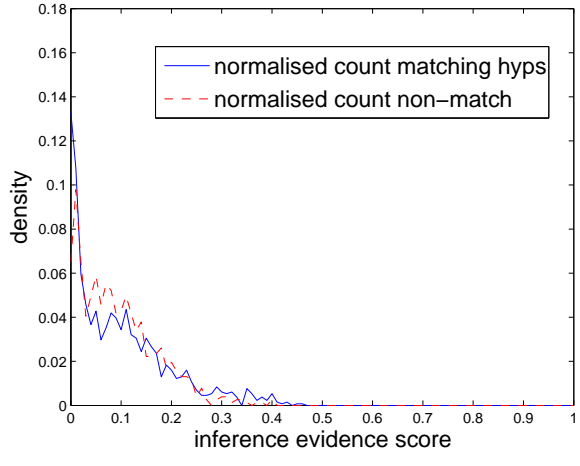
Our results support a general preference for IE scores when applying POMDP RL approaches to dialogue management. This provides independent support for preferring IE scores as was suggested by Thomson et al. (2008).

We have also sketched out how the data derived as part of the experimental analysis lends itself to improved training by providing a data drive simulated ASR-SLU noise model.

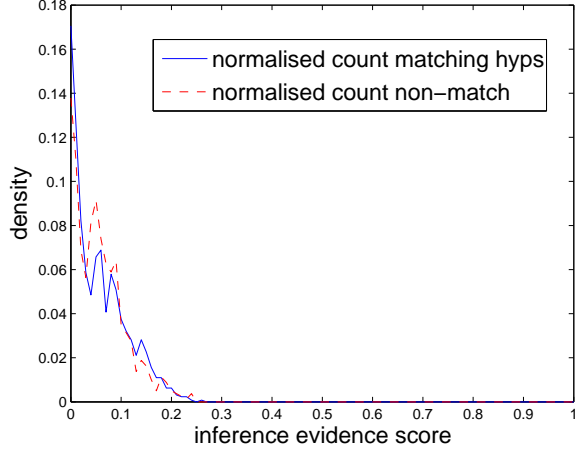
Finally, improved confidence scoring from automatic speech recognisers, as explored in this paper, is clearly an important ingredient of many future spoken dialogue systems.



(a) 1st hypothesis, IE score



(b) 2nd hypothesis, IE score



(c) 3rd hypothesis, IE score

Figure 2: Distribution of matching and non-matching parsed hypotheses against confidence score. All curves are normalised to give comparative probability distributions, *i.e.* the area under each curve sums to 1.0.

metric	hypothesis	gradient	intercept	MSE
IE	1st	0.5876	0.4382	0.0124
"	2nd	0.4270	0.5652	0.0083
"	3rd	0.3794	0.5896	0.0051
av. word conf.	1st	1.1403	-0.3416	0.0150
"	2nd	-0.2247	0.8022	0.0110
"	3rd	0.2449	0.3784	0.0143

Table 2: Intercept, gradient and mean squared error (MSE) for linear regressions on the data plotted in figure 1

hypothesis	$\sum_b S_{match}^{b,j,m}$	$\sum_b S_{nonmatch}^{b,j,m}$	freq. matching	freq. non-matching
1st	1,535	541	0.74	0.26
2nd	1,310	766	0.63	0.37
3rd	1,278	798	0.62	0.38

Table 3: Summed counts of matching and non-matching parsed hypotheses. This table gives the overall counts and frequency of each of the parsed hypotheses matching the parsed transcription summing over the range of confidence score values.

Acknowledgements

This work was funded by the EPSRC (project number EP/E019501/1).

References

- Malte Gabsdil and Oliver Lemon. 2004. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of ACL-04*, pages 344–351.
- M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, K. Yu, and S. Young. 2008. Training and evaluation of the HIS POMDP dialogue system in noise. In *Proceedings of SIGDial*.
- James Henderson and Oliver Lemon. 2008. Mixture Model POMDPs for Efficient Handling of Uncertainty in Dialogue Management. In *Proceedings of ACL*.
- R. Jonson. 2006. Dialogue Context-Based Re-ranking of ASR Hypotheses. In *Proceedings IEEE 2006 Workshop on Spoken Language Technology*.
- Oliver Lemon and Ioannis Konstas. 2009. User simulations for context-sensitive speech recognition in spoken dialogue systems. In *European Conference of the Association for Computational Linguistics*.
- Oliver Lemon, Kallirroi Georgila, and James Henderson. 2006. Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation. In *IEEE/ACL Spoken Language Technology*.
- Diane J. Litman, Julia Hirschberg, and Marc Swerts. 2000. Predicting Automatic Speech Recognition Performance Using Prosodic Cues. In *Proceedings of NAACL*.
- L Mangu, E Brill, and A Stolcke. 2000. Finding consensus among words: Lattice-based word error minimisation. *Computer Speech and Language*, 14(4):373–400.
- Ivan Meza-Ruiz, Sebastian Riedel, and Oliver Lemon. 2008. Accurate statistical spoken language understanding from limited development resources. In *ICASSP 08*.
- A. Ranta. 2004. Grammatical framework. a type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189.
- Blaise Thomson, Jost Schatzmann, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Training a real-world POMDP-based Dialog System. In *Proc. of Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 9–16. ACL.
- B. Thomson, K. Yu, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, and S. Young. 2008. Evaluating semantic-level confidence scores with multiple hypotheses. In *Proceedings of Interspeech*.
- Marilyn Walker, Jerry Wright, and Irene Langkilde. 2000. Using Natural Language Processing and Discourse Features to Identify Understanding Errors in a Spoken Dialogue System. In *Proceedings of ICML-2000*.
- SJ Young, J Schatzmann, K Weilhammer, and H Ye. 2007. The Hidden Information State Approach to Dialog Management. In *ICASSP 2007*.