Using Prosodic Information to Constrain Language Models for Spoken Dialogue

Paul Taylor, Hiroshi Shimodaira, Stephen Isard, Simon King and Jaqueline Kowtko

Centre for Speech Technology Research
University of Edinburgh 80 South Bridge, EH1 1HN
Paul.Taylor@ed.ac.uk

ABSTRACT

We present work intended to improve speech recognition performance for computer dialogue by taking into account the way that dialogue context and intonational tune interact to limit the possibilities for what an utterance might be. We report here on the extra constraint achieved in a bigram language model, expressed in terms of entropy, by using separate submodels for different sorts of dialogue acts, and trying to predict which submodel to apply by analysis of the intonation of the sentence being recognised.

1. INTRODUCTION

The ultimate goal of the work described here is to improve speech recognition performance for computer dialogue by taking into account the way that dialogue context and intonational tune interact to limit the possibilities for what an utterance might be.

For example, suppose that you ask a yes/no question, and receive a short reply uttered fairly low in the speaker's pitch range and without much movement in pitch. The chances are that the reply amounts to either "yes" or "no". It might be "sure" or "uh-uh" or some other equivalent, but a reply that denied an assumption in the question, asked for clarification, or simply refused to answer would probably be marked by a pitch accent. This is a simple case of the sort of constraint that we are exploiting.

We are developing our system on dialogues in the Map Task domain, using data from two English corpora, referred to as the HCRC and DCIEM Corpora [1] [2]. Our approach combines an intonational event finder based on the work described in Taylor [6] with a set of bigram language models corresponding to different dialogue acts [3].A neural net is trained to take the output of the event finder and predict the most likely dialogue act given that output.

Although this work is ultimately intended for application in a speech recognition system, we report here just on the theoretical constraint achieved in the language model, expressed in terms of reduction in perplexity.

Our basic idea is that the bigram language models for different individual dialogue acts should have lower entropy - more predictive power - than the general model for our data set as a whole. This means that we should be able to get improved performance if we were clairvoyant and knew in advance which model to use for a given utterance. We demonstrate a method for predicting which kind of game move a given utterance constitutes, by performing an analysis of the utterance's intonational tune. We investigate various combinations of this imperfect prediction with game move bigram models to see which give improved performance.

2. GAME MOVES IN DIALOGUES

Our dialogue analysis is based on the theory of conversational games first introduced by Power [4] and adapted for Maptask dialogues in [3]. Conversational games are conventional sequences of acts, such as question - answer - acknowledgement, or, indeed, request - non-linguistic-action - acknowledgement. The individual acts are termed moves. We distinguish 12 types of game moves, 6 of which initiate games and 6 which respond to and acknowledge earlier moves in the game.

Game move types correlate with syntax, but they are not syntactically defined because, for example, the move initiating a question game can have declarative syntax and indicate its question force intonationally. Game moves also differ from conventionally defined speech acts, because they are characterised by the purpose for which they are uttered. For instance, we distinguish between questions seeking new information and ones seeking confirmation of information that the speaker already believes to be true, and between statements which answer a conversational partner's question and ones which volunteer unsolicited information.

It is our hypothesis that the combination of game context and intonation can be used to predict game moves with high accuracy, but in this paper we investigate the predictive power of intonation on its own.

3. THE MAP TASK

In the Map Task [1] each of two participants has a schematic map which the other cannot see. The participants collaborate to reproduce on one of the maps a route already printed on the other. The two maps are not identical; the participants are told that they have been drawn by different explorers. Each map indicates about fifteen landmarks, and a landmark present on one map may be absent, or differently labelled, on the other. Although the participant with the pre-printed route is designated the Instruction Giver, and the other as the Instruction Follower, no restrictions are placed on what either can say.

The HCRC corpus consists of 128 dialogues between Glasgow University undergraduates. It contains about 150,000 words and has been transcribed at the word level and coded with a game move type for every utterance. The vocabulary size is 1810. The DCIEM corpus consists of 216 dialogues between Canadian military reservists. It has been transcribed at the word level, and the utterances of 4 dialogues (6460 words) were classified into game move types for the work described here.

4. INTONATION RECOGNITION AND TUNE ANALYSIS

Our game move prediction is based on the an analysis of the distinctive intonational pattern or *tune* of each utterance. We base our intonational tune recognition on the analysis of pitch accents and boundary rises (collectively known as intonational events).

Rather than using a set of discrete labels to describe intonational events (such as those used in the ToBI system [5]), we assign 4 continuous variables to each event. This has the advantage of avoiding any thresholding in this intermediate stage.

The values represent the 4 parameters in the Tilt intonation model [7]. These are: amplitude (size of F_0 excursion from immediate environment), duration (duration of event), position (absolute F_0 value at the start of the event) and tilt (representing the shape of the event). In the model, events are described as rises followed by falls, with the provision that one or the other may be absent. The tilt parameter is calculated by comparing the relative amplitudes and durations of the rise and fall components.

The approximate position and type of an utterance's intonational events are located by HMMs. These are trained on hand labelled examples of five types of units: **a** (pitch accent), **b** (boundary rise), **ab** (simultaneous occurrence of pitch accent and boundary rise), **c** (connection, the contour between accents) and silence. The HMMs work on feature vectors of 12 cepstral coefficients, smoothed F_0, delta smoothed F_0 and energy. The system correctly recognises 80% of intonational events[1].

After the approximate location of each event is determined by the HMM recogniser, a parameter fitting technique is used on the F_0 contour to determine the 4 tilt parameters for that event.

In common with much intonational analysis, we assume that the tune of a utterance is mainly characterised by the last pitch accent (the nucleus) and the boundary rise (if any) which follows it. Thus we characterise tune as an 8-component vector, the first 4 values representing the pitch accent and the second 4 the boundary rise. There are two special cases: a) when there is no event at the end of the utterance, it is assumed that no boundary rise exists and the 4 values are set to 0.0 b) where the nuclear accent and a boundary rise both occur at the end of the phrase, the second 4 values are the same as the first. The tune vectors are then normalised so that each parameter lies in the range -1 to +1.

A standard single hidden layer neural network, trained using back propagation, is used to predict game move from intonational tune. The network has 8 input nodes, one for each component of the tune vector, 20 hidden nodes and 12 output nodes, one for each game move.

As our speech recogniser operates on the DCIEM version of the corpus, we use only DCIEM data for intonation training and testing.

In an a speaker independent open test, the first choice of the neural network correctly predicts game moves 45% of the time. Although this is far from perfect, it is well above chance level with 12 game move types.

5. LANGUAGE MODELLING WITH MOVE SPECIFIC BIGRAMS

The main goal of a language model in speech recognition is to constrain the possibilities that the acoustic/phonetic component needs to consider, and thereby make it more likely to arrive at the correct answer. Bigram models are used in pursuit of the this goal as is common in most speech recognition systems. The power (i.e., degree of constraint) of a bigram can be expressed in terms of its entropy with respect to a data set. Entropy figures are sometimes transformed into perplexity as an aid to intuition, according to the formula $P = 2^{H(L)}$.

Perplexity gives an estimate of the average number of choices of following word that the system is faced with, given a decision for the current word. In our system we consider a bigram model for the whole corpus, as well as sub-bigrams for the sets of utterances assigned to each individual game move. The entire HCRC corpus has been game move coded by hand and as such allows for the training of more robust bigrams than the DCIEM data. Thus the results for bigrams

[1] The system was trained on multiple speakers but is not fully speaker-independent because the test speakers are the same as the training speakers. The actual test utterances are separate of course.

given here are trained and tested on material from the HCRC corpus alone.

The word entropy of a model with respect to a test set can be expressed as:

$$H(W) = -(1/K)\sum_i (1/n_i) log P(s_i)$$

where $P(s_i)$ is the probability assigned by the model to sentence s_i, n_i is the number of words in sentence i of the set, and K is the number of sentences in the test set. This can be rewritten

$$H(W) = -(1/K)\sum_i (1/n_i)\sum_j log P(w_{ij})$$

where and $P(w_{ij})$ is the probability assigned by the model to word j of sentence i. Interpreting P as giving bigram probabilities we can compute the entropies of the general model, and the individual sub-models for the various game move types (see table 1).

To combine the entropies of the individual game moves and the game move predictor we unpack P above as $P(s_i|m_i)P(m_i)$, where m_i is the correct game move class for s_i, $P(s_i|m_i)$ is the probability of sentence s_i in the bigram model for game move m_i, and $P(m_i)$ is the probability of the intonational predictor correctly identifying m_i for sentence s_i.

The formula can thus be rewritten as

$$H = (1/K)\sum_i (1/n_i) log P(s_i|m_i)P(m_i)$$

$$= (1/K)\sum_i (1/n_i)\sum_j log P(w_{ij}|m_i) + (1/n_i) log P(m_i)$$

Incorporating intonational prediction of game moves will give an improved - more constrained - language model if H as just defined is less than $H(W)$ above, when $P(w_{ij})$ is interpreted as the probability assigned by the general bigram model. Note that this formula incorporates the conservative assumption that we need to correctly identify m_i in order to assign a non-zero probability to s_i. Since sentences will in general have non-zero probabilities with respect to "wrong" models, use of the formula underestimates the power of the model.

6. RESULTS

Table 1 shows three measures of bigram statistics representing closed test, open test and a special open test calculation. The first point to note is that in most cases the closed test bigrams (H_c) have significantly lower entropies than the whole-task bigram. This is a demonstration that the division of utterances in this way does produce more tightly constrained language models. The open test entropies (H_o) follow roughly the same pattern as the closed test ones, in that

Game-Move	N	H_c	H_o	H_{os}
whole-task	26736	3.7495	4.2421	3.6153
acknowledge	5320	1.9498	2.9368	1.9568
align	1810	2.5793	3.8005	2.1957
check	2260	3.8764	6.1647	3.7942
clarify	1276	3.9308	6.7900	3.8028
explain	2183	3.9021	7.1133	3.7795
instruct	4379	4.1171	5.4649	3.9852
query-w	779	3.2685	5.9838	3.0803
query-yn	1785	3.2028	5.2361	3.0195
ready	1884	1.4693	1.6855	1.4021
reply-n	894	1.1002	2.0818	0.8179
reply-w	960	3.6196	7.0179	3.4419
reply-y	3206	1.9615	2.5912	1.8820

Table 1: Entropy measurements for the whole-task bigram and move-specific sub-bigrams. H_c is closed test, H_o is open test and H_{os} is the special case open test. N gives the number of utterances of that type in the whole corpus.

acknowledge has a relatively low entropy whereas **clarify** has a relatively high entropy. However, seven of the game moves have entropies which are much higher than the whole-task entropy. It is our opinion that this increase is due to insufficient training data for the sub-bigrams. We only have about 1000 examples of a game move such as **clarify** and this is probably too few to calculate a robust bigram for an 1800 word vocabulary. In recognition of this, we have calculated an additional measure H_{oc} which is an open test that discounts word-pair sequences which occur in the test set but not in the training set. Here we see that in all cases the sub-bigrams are either slightly higher or lower than the whole-task entropy. This figure cannot be used as a true measure of performance, but it is an indication that with more data (or more sophisticated bigram estimation) we should expect the move-specific bigram entropies to be lower in open test conditions.

Table 2 shows the overall entropy for four conditions. Condition 1 represents the whole-task bigram. Condition 2 represents the case where we combine intonational prediction with game move-specific bigrams. The results show slightly worse entropy that those for the whole-task bigram for the open test. However the special open test figure shows that the entropy is much lower than the whole-task bigram.

In light of the apparent sparse data problem, we reduced the number of sub-bigrams. From examination of the transcriptions, game moves were clustered into five categories that were deemed to have similar syntactic properties. Condition 3 represents this clustering approach. These were (acknowledge align ready), (check query-yn), (clarify explain instruct), (query-w reply-w) and (reply-n reply-y). As well as giving more training data for bigram estimation, the reduction in number of classes also helps the performance of the game-move predictor, which gives a 55% recognition rate for this clustering. Table 2 shows that both the genuine open test and the special open test give lower entropies than the whole-task bigram.

Condition	% A	H_o	H_{os}	PP_o	PP_{os}
1	-	4.2421	3.6153	18.92	12.26
2	45%	4.3624	2.5498	20.57	5.86
3	55%	4.0872	2.9780	17.00	7.88
4	-	5.3621	3.2653	24.66	10.25

Table 2: Results for conditions 1) whole-task bigram; 2) 12 game move bigrams; 3) clustered bigrams; 4) randomly partitioned bigrams. %A is the accuracy of the game move predictor, H_o is open test, H_{os} is the special case open test and PP is the equivalent perplexity measure.

Condition 4 is included for interest and shows the performance when the data is arbitrarily divided into 12 sub groups, for which we assume the same prediction power as for condition 2. It can be clearly seen that the performance is much worse. This is evidence that the game move analysis is meaningful and is a sensible way to divide a corpus.

7. DISCUSSION

The results indicate that in principle, dividing a task in this fashion can produce a more tightly constrained overall model. We have identified two main factors within the current framework which will improve overall performance: game move prediction accuracy and amount of training data.

To investigate the effect of game move prediction accuracy we have made estimates of what overall entropy scores we can expect for a given accuracy. It is impossible to calculate exactly what entropy a particular prediction accuracy will produce as the total entropy depends on what sort of errors are made by the predictor. However, for a particular move predictor of average accuracy 75%, we found that the open test entropy was slightly lower in the 12 game move case. Thus we can start expecting overall improvements in genuine open tests when prediction accuracy approaches this figure.

In the 12 game move case, the bigram estimations are too sparse. We have as yet not employed any sophisticated smoothing techniques in bigram calculation, and this along with more data should make the open test results better. The reduction in sub-bigrams caused by clustering produces both improved game move prediction and better bigram estimation, which is the reason for the improved performance in this case.

Our most important finding is that using game move specific bigrams reduces the perplexity of a language model. This is only useful if we can predict which game move an utterance belongs to prior to word recognition. Our work so far has concentrated on using intonational tune for prediction but the general technique still holds for any method of move prediction.

An obvious next step is to use dialogue context. Dialogues follow patterns in that one participant's choice of utterance is partly dependent on the other participant's previous utterance. Furthermore, we believe there to be an interaction between intonation and context. In particular, a "default" following move type at any stage is less likely to be intonationally marked, as in the example given above in the introduction.

It is clear from table 1 that most game move bigrams have entropies which are significantly less than that for the whole task. This is an indication of the uniformity of the utterances assigned to that class. However, the technique we have developed here will work in principle for any division of a corpus into utterance types. The success of a grouping depends on two factors: that the sub-bigrams have a low entropy and that the types can be predicted, either acoustically or from context.

Notes: The DCIEM and HCRC maptask corpora are publicly available databases distributed on CD-ROM. The basic speech data and transcripts are distributed by DCIEM, HCRC at the University of Edinburgh, and by the Linguistic Data Consortium. The intonation labels and game move coding are not distributed with the CDs but can be obtained from the authors.

We gratefully acknowledge the support of EPSRC in funding this work through grant number GR/J55106.

8. REFERENCES

1. Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth H. Boyle, Gwyneth M. Doherty, Simon C. Garrod, Stephen D. Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. The hcrc map task corpus. *Language and Speech*, 34(4):351–366, 1991.
2. Ellen G. Bard, Catherine Sotillo, Anne H. Anderson, and M.M. Taylor. The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In *Proc. of the ESCA-NATO Tutorial and Workshop on Speech under Stress, Lisbon*, 1995.
3. Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The coding of dialogue structure in a corpus. In J.A. Andernach, S.P. van de Burgt, and G.F. van der Hoeven, editors, *Proceedings of the Ninth Twente Workshop on Language Technology: Corpus-based Approaches to Dialogue Modelling*. Universiteit Twente, Enschede, 1995.
4. R. Power. The organization of purposeful dialogues. *Linguistics*, 17:107–152, 1979.
5. K. Silverman, J. Piterelli, M. Beckman, J. Pierrehumbert, R. Ladd, C. Wightman, M. Ostendorf, and J. Hirschberg. Tones and break indices: a standard for prosodic transcription. In *International Conference on Speech and Language Processing '92, Banff, Canada*, 1992.
6. Paul A. Taylor. Using neural networks to locate pitch accents. In *Proc. Eurospeech '95, Madrid*, 1995.
7. Paul A. Taylor and Alan W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Second ESCA/IEEE Workshop on Speech Synthesis, New York, U.S.A*, 1994.