User Simulations for context-sensitive speech recognition in Spoken Dialogue Systems

Oliver Lemon
Edinburgh University
olemon@inf.ed.ac.uk

Ioannis Konstas
University of Glasgow
konstas@dcs.gla.ac.uk

Abstract

We use a machine learner trained on a combination of acoustic and contextual features to predict the accuracy of incoming n-best automatic speech recognition (ASR) hypotheses to a spoken dialogue system (SDS). Our novel approach is to use a simple statistical User Simulation (US) for this task, which measures the likelihood that the user would say each hypothesis in the current context. Such US models are now common in machine learning approaches to SDS, are trained on real dialogue data, and are related to theories of "alignment" in psycholinguistics. We use a US to predict the user's next dialogue move and thereby re-rank n-best hypotheses of a speech recognizer for a corpus of 2564 user utterances. The method achieved a significant relative reduction of Word Error Rate (WER) of 5% (this is 44% of the possible WER improvement on this data), and 62% of the possible semantic improvement (Dialogue Move Accuracy), compared to the baseline policy of selecting the topmost ASR hypothesis. The majority of the improvement is attributable to the User Simulation feature, as shown by Information Gain analysis.

1 Introduction

A crucial problem in the design of spoken dialogue systems (SDS) is to decide for incoming recognition hypotheses whether a system should *accept* (consider correctly recognized), *reject* (assume misrecognition), or *ignore* (classify as noise or speech not directed to the system) them.

Obviously, incorrect decisions at this point can have serious negative effects on system usability and user satisfaction. On the one hand, accepting misrecognized hypotheses leads to misunderstandings and unintended system behaviors which are usually difficult to recover from. On the other hand, users might get frustrated with a system that behaves too cautiously and rejects or ignores too many utterances. Thus an important feature in dialogue system engineering is the tradeoff between avoiding task failure (due to misrecognitions) and promoting overall dialogue efficiency, flow, and naturalness.

In this paper, we investigate the use of machine learning trained on a combination of acoustic features and features computed from dialogue context to predict the quality of incoming n-best recognition hypotheses to a SDS. These predictions are then used to select a "best" hypothesis and to decide on appropriate system reactions. We evaluate this approach in comparison with a baseline system that works in the standard way: always choosing the topmost hypothesis in the n-best list. In such systems, complex repair strategies are required when the top hypothesis is incorrect.

The main novelty of this work is that we explore the use of predictions from simple statistical User Simulations to re-rank n-best lists of ASR hypotheses. These User Simulations are now commonly used in statistical learning approaches to dialogue management (Williams and Young, 2003; Schatzmann et al., 2006; Young, 2006; Young et al., 2007; Schatzmann et al., 2007), but they have not been used for context-sensitive ASR before.

In our model, the system's "belief" $b(h)$ in a recognition hypothesis h is factored in two parts: the observation probability $P(o|h)$ (approximated by the ASR confidence score) and the User Simulation probability $P(h|us, C)$ of the hypothesis:

$$b(h) = P(o|h).P(h|us, C) \quad (1)$$

where us is the state of the User Simulation in context C. The context is simply a window of di-

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 505–513,
Athens, Greece, 30 March – 3 April 2009.

alogue acts in the dialogue history, that the US is sensitive to (see section 3).

The paper is organized as follows. After a short relation to previous work, we describe the data (Section 5) and derive baseline results (Section 6). Section 3 describes the User Simulations that we use for re-ranking hypotheses. Section 7 describes our learning experiments for classifying and selecting from n-best recognition hypotheses and Section 9 reports our results.

2 Relation to Previous Work

In psycholinguistics, the idea that human dialogue participants simulate each other to some extent is gaining currency. (Pickering and Garrod, 2007) write:

> "if B overtly imitates A, then A's comprehension of B's utterance is facilitated by A's memory for A's previous utterance."

We explore aspects of this idea in a computational manner. Similar work in the area of spoken dialogue systems is described below.

(Litman et al., 2000) use acoustic-prosodic information extracted from speech waveforms, together with information derived from their speech recognizer, to automatically predict misrecognized turns in a corpus of train-timetable information dialogues. In our experiments, we also use recognizer confidence scores and a limited number of acoustic-prosodic features (e.g. amplitude in the speech signal) for hypothesis classification, but we also use User Simulation predictions.

(Walker et al., 2000) use a combination of features from the speech recognizer, natural language understanding, and dialogue manager/discourse history to classify hypotheses as correct, partially correct, or misrecognized. Our work is related to these experiments in that we also combine confidence scores and higher-level features for classification. 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. We go a step further in that we classify n-best hypotheses and then select among the alternatives. We also explore the use of more dialogue and task-oriented features (e.g. the dialogue move type of a recognition hypothesis) for classification.

(Gabsdil and Lemon, 2004) similarly perform reordering 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.

(Jonson, 2006) classifies recognition hypotheses with labels denoting acceptance, clarification, confirmation and rejection. These labels were learned 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.

As well as the use of a User Simulation, the main difference between our approach and work on hypothesis reordering (e.g. (Chotimongkol and Rudnicky, 2001)) is that we make a decision regarding whether a dialogue system should accept, clarify, reject, or ignore a user utterance. Like (Gabsdil and Lemon, 2004; Jonson, 2006), our approach is more generally applicable than preceding research, since we frame our methodology in the *Information State Update* (ISU) approach to dialogue management (Traum et al., 1999) and therefore expect it to be applicable to a range of related multimodal dialogue systems.

3 User Simulations

What makes this study different from the previous work in the area of post-processing of the ASR hypotheses is the incorporation of a User Simulation output as an additional feature. The history of a dialogue between a user and a dialogue system plays an important role as to what the user might be expected to say next. As a result, most of the studies mentioned in the previous section make various efforts to capture history by including relevant features directly in their classifiers.

Various statistical User Simulations have been trained on corpora of dialogue data in order to simulate real user behaviour (Schatzmann et al., 2006; Young, 2006; Georgila et al., 2006; Young et al., 2007; Schatzmann et al., 2007). We developed a simple n-gram User Simulation, using n-grams of dialogue moves. It treats a dialogue as a sequence of lists of consecutive user and system turns in a high level semantic representation, i.e.

$< SpeechAct >, < Task >$ pairs, for example $< provide\_info >, < music\_genre(punk) >$. It takes as input the $n-1$ most recent lists of $< SpeechAct >, < Task >$ pairs in the dialogue history, and uses the statistics in the training set to compute a distribution over the possible next user actions. If no n-grams match the current history, the model can back-off to n-grams of lower order. We use this model to assess the likelihood of each candidate ASR hypothesis. Intuitively, this is the likelihood that the user really would say the hypothesis in the current dialogue situation. The benefit of using n-gram models is that they are fast and simple to train even on large corpora.

The main hypothesis that we investigate is that by using the User Simulation model to predict the next user utterance, we can effectively increase the performance of the speech recogniser module.

4 Evaluation metrics

To evaluate performance we use Dialogue Move Accuracy (DMA), a strict variant of Concept Error Rate (CER) as defined by (Boros et al., 1996), which takes into account the semantic aspects of the difference between the classified utterance and the true transcription. CER is similar to WER, since it takes into account deletions, insertions and substitutions on the semantic (rather than the word) level of the utterance. DMA is stricter than CER in the sense that it does not allow for partial matches in the semantic representation. In other words, if the classified utterance corresponds to the same semantic representation as the transcribed then we have 100% DMA, otherwise 0%.

Sentence Accuracy (SA) is the alignment of a single hypothesis in the n-best list with the true transcription. Similarly to DMA, it accounts for perfect alignment between the hypothesis and the transcription, i.e. if they match perfectly we have 100% SA, otherwise 0%.

5 Data Collection

For our experiments, we use data collected in a user study with the Town-Info spoken dialogue system, using the HTK speech recognizer (Young, 2007). In this study 18 subjects had to solve 10 search/browsing tasks with the system, resulting in 180 complete dialogues and 2564 utterances (average 14.24 user utterances per dialogue).

For each utterance we have a series of files of 60-best lists produced by the speech recogniser, namely the transcription hypotheses on a sentence level along with the acoustic model score and the equivalent transcriptions on a word level, with information such as the duration of each recognised frame and the confidence score of the acoustic and language model of each word.

5.1 Labeling

We transcribed all user utterances and parsed the transcriptions offline using a natural language understanding component (a robust Keyword Parser) in order to get a gold-standard labeling of the data.

We devised four labels with decreasing order of confidence: 'opt' (optimal), 'pos' (positive), 'neg' (negative), 'ign' (ignore). These are automatically generated using two different modules: a keyword parser that computes the $< SpeechAct >< Task >$ pair as described in the previous section and a Levenshtein Distance calculator, for the computation of the DMA and WER of each hypothesis respectively. The reason for opting for a more abstract level, namely the semantics of the hypotheses rather than individual word recognition, is that in SDS it is usually sufficient to rely on the meaning of message that is being conveyed by the user rather than the precise words that they used.

Similar to (Gabsdil and Lemon, 2004; Jonson, 2006) we ascribe to each utterance either of the 'opt', 'pos', 'neg', 'ign' labels according to the following schema:

- **opt**: The hypothesis is perfectly aligned and semantically identical to the transcription
- **pos**: The hypothesis is not entirely aligned (WER < 50) but is semantically identical to the transcription
- **neg**: The hypothesis is semantically identical to the transcription but does not align well (WER > 50) or is semantically different to the transcription
- **ign**: The hypothesis was not addressed to the system (crosstalk), or the user laughed, coughed, etc.

The 50% value for the WER as a threshold for the distinction between the 'pos' and 'neg' category is adopted from (Gabsdil, 2003), based on the fact that WER is affected by concept accuracy (Boros et al., 1996). In other words, if a hypothesis is erroneous as far as its transcript is concerned

Transcript: I'd like to find a bar please	
I WOULD LIKE TO FIND A BAR PLEASE	pos
I LIKE TO FIND A FOUR PLEASE	neg
I'D LIKE TO FIND A BAR PLEASE	opt
WOULD LIKE TO FIND THE OR PLEASE	ign

Table 1: Example hypothesis labelling

then it is highly likely that it does not convey the correct message from a semantic point of view. We always label conceptually equivalent hypotheses to a particular transcription as potential candidate dialogue strategy moves, and total misrecognitions as rejections. In table 5.1 we show examples of the four labels. Note that in the case of silence, we give an 'opt' to the empty hypothesis.

6 The Baseline and Oracle Systems

The baseline for our experiments is the behavior of the Town-Info spoken dialogue system that was used to collect the experimental data. We evaluate the performance of the baseline system by analyzing the dialogue logs from the user study.

As an oracle for the system we defined the choice of either the first 'opt' in the n-best list, or if this does not exist the first 'pos' in the list. In this way it is guaranteed that we always get as output a perfect match to the true transcript as far as its Dialogue Move is concerned, provided there exists a perfect match somewhere in the list.

6.1 Baseline and Oracle Results

Table 2 summarizes the evaluation of the baseline and oracle systems. We note that the Baseline system already performs quite well on this data, when we consider that in about 20% of n-best lists there is no semantically correct hypothesis.

	Baseline	Oracle
WER	47.72%	42.16%
DMA	75.05%	80.20%
SA	40.48%	45.27%

Table 2: Baseline and Oracle results (statistically significant at $p < 0.001$)

7 Classifying and Selecting N-best Recognition Hypotheses

We use a threshold (50%) on a hypothesis' WER as an indicator for whether hypotheses should be clarified or rejected. This is adopted from (Gabsdil, 2003), based on the fact that WER correlates with concept accuracy (CA, (Boros et al., 1996)).

7.1 Classification: Feature Groups

We represent recognition hypotheses as 13-dimensional feature vectors for automatic classification. The feature vectors combine recognizer confidence scores, low-level acoustic information, and information from the User Simulation.

All the features used by the system are extracted by the dialogue logs, the n-best lists per utterance and per word and the audio files. The majority of the features chosen are based on their success in previous systems as described in the literature (see section 2). The novel feature here is the User Simulation score which may make redundant most of the dialogue features used in other studies.

In order to measure the usefulness of each candidate feature and thus choose the most important we use the metrics of Information Gain and Gain Ratio (see table 3 in section 8.1) on the whole training set, i.e. 93240 hypotheses.

In total 13 attributes were extracted, that can be grouped into 4 main categories; those that concern the current hypothesis to be classified, those that concern low-level statistics of the audio files, those that concern the whole n-best list, and finally the User Simulation feature.

- Current Hypothesis Features (CHF) (6): acoustic score, overall model confidence score, minimum word confidence score, grammar parsability, hypothesis length and hypothesis duration.
- Acoustic Features (AF) (3): minimum, maximum and RMS amplitude
- List Features (LF) (3): n-best rank, deviation of confidence scores in the list, match with most frequent Dialogue Move
- User Simulation (US) (1): User Simulation confidence score

The **Current Hypothesis features (CHF)** were extracted from the n-best list files that contained the hypotheses' transcription along with overall acoustic score per utterance and from the equivalent files that contained the transcription of each word along with the start of frame, end of frame and confidence score:

Acoustic score is the negative log likelihood ascribed by the speech recogniser to the whole hypothesis, being the sum of the individual word acoustic scores. Intuitively this is considered to be helpful since it depicts the confidence of the statistical model only for each word and is also adopted in previous studies. Incorrect alignments shall tend to adapt less well to the model and thus have low log likelihood.

Overall model confidence score is the average of the individual word confidence scores.

Minimum word confidence score is also computed by the individual word transcriptions and accounts for the confidence score of the word which the speech recogniser is least certain of. It is expected to help our classifier distinguish between poor overall hypothesis recognitions since a high overall confidence score can sometimes be misleading.

Grammar Parsability is the negative log likelihood of the transcript for the current hypothesis as produced by the Stanford Parser, a wide-coverage Probabilistic Context-Free Grammar (PCFG) (Klein and Manning, 2003) [1]. This feature seems helpful since we expect that a highly ungrammatical hypothesis is likely not to match with the true transcription semantically.

Hypothesis duration is the length of the hypothesis in milliseconds as extracted from the n-best list files with transcriptions per word that include the start and the end time of the recognised frame. The reason for the inclusion of this feature is that it can help distinguish between short utterances such as yes/no answers, medium-sized utterances of normal answers and long utterances caused by crosstalk.

Hypothesis length is the number of words in a hypothesis and is considered to help in a similar way as the above feature.

The **Acoustic Features (AF)** were extracted directly from the wave files using SoX: Minimum, maximum and RMS amplitude are straightforward features common in the previous studies mentioned in section 2.

The **List Features (LF)** were calculated based on the n-best list files with transcriptions per utterance and per word and take into account the whole list:

N-best rank is the position of the hypothesis in the list and could be useful in the sense that 'opt' and 'pos' may be found in the upper part of the list rather than the bottom.

Deviation of confidence scores in the list is the deviation of the overall model confidence score of the hypothesis from the mean confidence score in the list. This feature is extracted in the hope that it will indicate potential clusters of confidence scores in particular positions in the list, i.e. group hypotheses that deviate in a specific fashion from the mean and thus indicating them being classified with the same label.

Match with most frequent Dialogue Move is the only boolean feature and indicates whether the Dialogue Move of the current hypothesis, i.e. the pair of $< SpeechAct >< Task >$ coincides with the most frequent one. The trend in n-best lists is to have a majority of utterances that belong to one or two labels and only one hypothesis belonging to the 'opt' category and/or a few to the 'pos' category. As a result, the idea behind this feature is to extract such potential outliers which are the desired goal for the re-ranker.

Finally, the **User Simulation score** is given as an output from the User Simulation model and adapted for the purposes of this study (see section 3 for more details). The model is operating with 5-grams. Its input is given by two different sources: the history of the dialogue, namely the 4 previous Dialogue Moves, is taken from the dialogue log and the current hypothesis' semantic parse which is generated on the fly by the same keyword parser used in the automatic labelling.

User Simulation score is the probability that the current hypothesis' Dialogue Move has really been said by the user given the 4 previous Dialogue Moves. The potential advantages of this feature have been discussed in section 3.

7.2 Learner and Selection Procedure

We use the memory based learner TiMBL (Daelemans et al., 2002) to predict the class of each of the 60-best recognition hypotheses for a given utterance.

TiMBL was trained using different parameter combinations mainly choosing between number of k-nearest neighbours (1 to 5) and distance metrics (Weighted Overlap and Modified Value Difference Metric). In a second step, we decide which (if any) of the classified hypotheses we actually want to pick as the best result and how the user utterance should be classified as a whole.

[1] http://nlp.stanford.edu/software/lex-parser.shtml

1. Scan the list of classified n-best recognition hypotheses top-down. Return the first result that is classified as 'opt'.

2. If 1. fails, scan the list of classified n-best recognition hypotheses top-down. Return the first result that is classified as 'pos'.

3. If 2. fails, count the number of negs and igns in the classified recognition hypotheses. If the number of negs is larger or equal than the number of igns then return the first 'neg'.

4. Else return the first 'ign' utterance.

8 Experiments

Experiments were conducted in two layers: the first layer concerns only the classifier, i.e. the ability of the system to correctly classify each hypothesis to either of the four labels 'opt', 'pos', 'neg', 'ign' and the second layer the re-ranker, i.e. the ability of the system to boost the speech recogniser's accuracy.

All results are drawn from the TiMBL classifier trained with the Weighted Overlap metric and k = 1 nearest neighbours settings. Both layers are trained on 75% of the same Town-Info Corpus of 126 dialogues containing 60-best lists for 1554 user utterances or a total of 93240 hypotheses. The first layer was tested against a separate Town-Info Corpus of 58 dialogues containing 510 user utterances or a total of 30600 hypotheses, while the second was tested on the whole training set with 10-fold cross-validation.

Using this corpus, a series of experiments was carried out using different sets of features in order to both determine and illustrate the increasing performance of the classifier. These sets were determined not only by the literature but also by the Information Gain measures that were calculated on the training set using WEKA, as shown in table 3.

8.1 Information Gain

Quite surprisingly, we note that the rank given by the Information Gain measure coincides perfectly with the logical grouping of the attributes that was initially performed (see table 3).

As a result, we chose to use this grouping for the final 4 feature sets on which the classifier experiments were performed, in the following order:
Experiment 1: List Features (LF)

InfoGain	Attribute
1.0324	userSimulationScore
0.9038	rmsAmp
0.8280	minAmp
0.8087	maxAmp
0.4861	parsability
0.3975	acousScore
0.3773	hypothesisDuration
0.2545	hypothesisLength
0.1627	avgConfScore
0.1085	minWordConfidence
0.0511	nBestRank
0.0447	standardDeviation
0.0408	matchesFrequentDM

Table 3: Information Gain

Experiment 2: List Features + Current Hypothesis Features (LF+CHF)
Experiment 3: List Features + Current Hypothesis Features + Acoustic Features (LF+CHF+AF)
Experiment 4: List Features + Current Hypothesis Features + Acoustic Features + User Simulation (LF+CHF+AF+US)

Note that the User Simulation score is a very strong feature, scoring first in the Information Gain rank, validating our central hypothesis.

The testing of the classifier using each of the above feature sets was performed on the remaining 25% of the Town-Info corpus comprising of 58 dialogues, consisting of 510 utterances and taking the 60-best lists resulting in a total of 30600 vectors. In each experiment we measured Precision, Recall, F-measure per class and total Accuracy of the classifier .

For the second layer, we used a trained instance of the TiMBL classifier on the 4th feature set (List Features + Current Hypothesis Features + Acoustic Features + User Simulation) and performed reranking using the algorithm presented in section 7.2 on the same training set used in the first layer using 10-fold cross validation.

9 Results and Evaluation

We performed two series of experiments in two layers: the first corresponds to the training of the classifier alone and the second to the system as a whole measuring the re-ranker's output.

Feature set (opt)	Precision	Recall	F1
LF	42.5%	58.4%	49.2%
LF+CHF	62.4%	65.7%	64.0%
LF+CHF+AF	55.6%	61.6%	58.4%
LF+CHF+AF+US	70.5%	73.7%	72.1%

Table 4: Results for the 'opt' category

Feature set (pos)	Precision	Recall	F1
LF	25.2%	1.7%	3.2%
LF+CHF	51.2%	57.4%	54.1%
LF+CHF+AF	51.5%	54.6%	53.0%
LF+CHF+AF+US	64.8%	61.8%	63.3%

Table 5: Results for the 'pos' category

9.1 First Layer: Classifier Experiments

In these series of experiments we measure precision, recall and F1-measure for each of the four labels and overall F1-measure and accuracy of the classifier. In order to have a better view of the classifier's performance we have also included the confusion matrix for the final experiment with all 13 attributes. Tables 4 -7 show per class and per attribute set measures, while Table 8 shows a collective view of the results for the four sets of attributes and the baseline being the majority class label 'neg'. Table 9 shows the confusion matrix for the final experiment.

In tables 4 - 8 we generally notice an increase in precision, recall and F1-measure as we progressively add more attributes to the system with the exception of the addition of the Acoustic Features which seem to impair the classifier's performance. We also make note of the fact that in the case of the 4th attribute set the classifier can distinguish very well the 'neg' and 'ign' categories with 86.3% and 99.9% F1-measure respectively. Most importantly, we observe a remarkable boost in F1-measure and accuracy with the addition of the User Simulation score. We find a 37.36% relative increase in F1-measure and 34.02% increase

Feature set (neg)	Precision	Recall	F1
LF	54.2%	96.4%	69.4%
LF+CHF	70.7%	75.0%	72.8%
LF+CHF+AF	69.5%	73.4%	71.4%
LF+CHF+AF+US	85.6%	87.0%	86.3%

Table 6: Results for the 'neg' category

Feature set (ign)	Precision	Recall	F1
LF	19.6%	1.3%	2.5%
LF+CHF	63.5%	48.7%	55.2%
LF+CHF+AF	59.3%	48.9%	53.6%
LF+CHF+AF+US	99.9%	99.9%	99.9%

Table 7: Results for the 'ign' category

Feature set	F1	Accuracy
Baseline	-	51.1%
LF	37.3%	53.1%
LF+CHF	64.1%	64.8%
LF+CHF+AF	62.6%	63.4%
LF+CHF+AF+US	86.0%	84.9%

Table 8: F1-Measure and Accuracy for the four attribute sets

in the accuracy compared to the 3rd experiment, which contains all but the User Simulation score attribute and a 66.20% relative increase of the accuracy compared to the Baseline. In table 7 we make note of a rather low recall measure for the 'ign' category in the case of the LF experiment, suggesting that the list features do not add extra value to the classifier, partially validating the Information Gain measure (Table 3).

Taking a closer look at the 4th experiment with all 13 features we notice in table 9 that most errors occur between the 'pos' and 'neg' category. In fact, for the 'neg' category the False Positive Rate (FPR) is 18.17% and for the 'pos' 8.9%, all in all a lot larger than for the other categories.

9.2 Second Layer: Re-ranker Experiments

In these experiments we measure WER, DMA and SA for the system as a whole. In order to make sure that the improvement noted was really attributed to the classifier we computed the p-values for each of these measures using the Wilcoxon signed rank test for WER and McNemar chi-square test for the DMA and SA measures.

In table 10 we note that the classifier scores

	opt	pos	neg	ign
opt	232	37	46	0
pos	47	4405	2682	8
neg	45	2045	13498	0
ign	5	0	0	7550

Table 9: Confusion Matrix for LF+CHF+AF+US

	Baseline	Classifier	Oracle
WER	47.72%	45.27% **	42.16%***
DMA	75.05%	78.22% *	80.20% ***
SA	40.48%	42.26%	45.27%***

Table 10: Baseline, Classifier, and Oracle results (*** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$)

Label	Precision	Recall	F1
opt	74.0%	64.1%	68.7%
pos	76.3%	46.2%	57.6%
neg	81.9%	94.4%	87.7%
ign	99.9%	99.9%	99.9%

Table 11: Precision, Recall and F1: high-level features

45.27% WER making a notable relative reduction of 5.13% compared to the baseline and 78.22% DMA incurring a relative improvement of 4.22%. The classifier scored 42.26% on SA but it was not considered significant compared to the baseline ($0.05 < p < 0.10$). Comparing the classifier's performance with the Oracle it achieves a 44.06% of the possible WER improvement on this data, 61.55% for the DMA measure and 37.16% for the SA measure.

Finally, we also notice that the Oracle has a 80.20% for the DMA, which means that 19.80% of the n-best lists did not include at all a hypothesis that matched semantically to the true transcript.

10 Experiment with high-level features

We trained a Memory Based Classifier based only on the higher level features of merely the User Simulation score and the Grammar Parsability (US + GP). The idea behind this choice is to try and find a combination of features that ignores low level characteristics of the user's utterances as well as features that heavily rely on the speech recogniser and thus by default are not considered to be very trustworthy.

Quite surprisingly, the results taken from an experiment with just the User Simulation score and the Grammar Parsability are very promising and comparable with those acquired from the 4th experiment with all 13 attributes. Table 11 shows the precision, recall and F1-measure per label and table 12 illustrates the classifier's performance in comparison with the 4th experiment.

Table 12 shows that there is a somewhat considerable decrease in the recall and a corresponding increase in the precision of the 'pos' and 'opt' categories compared to the LF + CHF + AF + US attribute set, which account for lower F1-measures. However, all in all the US + GP set manages to classify correctly 207 more vectors and quite interestingly commits far fewer ties and manages to resolve more compared to the full 13 attribute set.

Feature set	F1	Accuracy	Ties
LF+CHF+AF+US	86.0%	84.9%	4993
US+GP	85.7%	85.6%	115

Table 12: F1, Accuracy and number of ties correctly resolved for LF+CHF+AF+US and US+GP feature sets

11 Conclusion

We used a combination of acoustic features and features computed from dialogue context to predict the quality of incoming recognition hypotheses to an SDS. In particular we use a score computed from a simple statistical User Simulation, which measures the likelihood that the user really said each hypothesis. The approach is novel in combining User Simulations, machine learning, and n-best processing for spoken dialogue systems. We employed a User Simulation model, trained on real dialogue data, to predict the user's next dialogue move. This prediction was used to re-rank n-best hypotheses of a speech recognizer for a corpus of 2564 user utterances. The results, obtained using TiMBL and an n-gram User Simulation, show a significant relative reduction of Word Error Rate of 5% (this is 44% of the possible WER improvement on this data), and 62% of the possible Dialogue Move Accuracy improvement, compared to the baseline policy of selecting the topmost ASR hypothesis. The majority of the improvement is attributable to the User Simulation feature. Clearly, this improvement would result in better dialogue system performance overall.

Acknowledgments

We thank Helen Hastie and Kallirroi Georgila. The research leading to these results has received funding from the EPSRC (project no. EP/E019501/1) and from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (CLASSiC project `www.classic-project.org`)

References

M. Boros, W. Eckert, F. Gallwitz, G. Görz, G. Hanrieder, and H. Niemann. 1996. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings ICSLP '96*, volume 2, pages 1009–1012, Philadelphia, PA.

Ananlada Chotimongkol and Alexander I. Rudnicky. 2001. N-best Speech Hypotheses Reordering Using Linear Regression. In *Proceedings of EuroSpeech 2001*, pages 1829–1832.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. TIMBL: Tilburg Memory Based Learner, version 4.2, Reference Guide. In *ILK Technical Report 02-01*.

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.

Malte Gabsdil. 2003. Classifying Recognition Results for Spoken Dialogue Systems. In *Proceedings of the Student Research Workshop at ACL-03*.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Proceedings of Interspeech/ICSLP*, pages 1065–1068.

R. Jonson. 2006. Dialogue Context-Based Re-ranking of ASR Hypotheses. In *Proceedings IEEE 2006 Workshop on Spoken Language Technology*.

D. Klein and C. Manning. 2003. Fast exact inference with a factored model for natural language parsing. *Journal of Advances in Neural Information Processing Systems*, 15(2).

Diane J. Litman, Julia Hirschberg, and Marc Swerts. 2000. Predicting Automatic Speech Recognition Performance Using Prosodic Cues. In *Proceedings of NAACL*.

M. Pickering and S. Garrod. 2007. Do people use language production to make predictions during comprehension? *Journal of Trends in Cognitive Sciences*, 11(3).

J Schatzmann, K Weilhammer, M N Stuttle, and S J Young. 2006. A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies. *Knowledge Engineering Review*, 21:97–126.

J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young. 2007. Agenda-based User Simulation for Bootstrapping a POMDP Dialogue System. In *Proceedings of HLT/NAACL*.

David Traum, Johan Bos, Robin Cooper, Staffan Larsson, Ian Lewin, Colin Matheson, and Massimo Poesio. 1999. A Model of Dialogue Moves and Information State Revision. Technical Report D2.1, Trindi Project.

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*.

Jason Williams and Steve Young. 2003. Using wizard-of-oz simulations to bootstrap reinforcement-learning-based dialog management systems. In *Proc. 4th SIGdial workshop*.

SJ Young, J Schatzmann, K Weilhammer, and H Ye. 2007. The Hidden Information State Approach to Dialog Management. In *ICASSP 2007*.

SJ Young. 2006. Using POMDPs for Dialog Management. In *IEEE/ACL Workshop on Spoken Language Technology (SLT 2006)*, Aruba.

Steve Young. 2007. ATK: An Application Toolkit for HTK, Version 1.6. Technical report, Cambridge University Engineering Department.