Combining Acoustic Confidences and Pragmatic Plausibility for Classifying Spoken Chess Move Instructions

Malte Gabsdil
Department of Computational Linguistics
Saarland University
Germany
gabsdil@coli.uni-sb.de

Abstract

This paper describes a machine learning approach to classifying n-best speech recognition hypotheses as either correctly or incorrectly recognised. The learners are trained on a combination of acoustic confidence features and move evaluation scores in a chess-playing scenario. The results show significant improvements over sharp baselines that use confidence rejection thresholds for classification.

1 Introduction

An important task in designing spoken dialogue systems is to decide whether a system should *accept* (consider correctly recognised) or *reject* (assume misrecognition) a user utterance. This decision is often based on acoustic confidence scores computed by the speech recogniser and a fixed confidence rejection threshold. However, a drawback of this approach is that it does not take into account that in particular dialogue situations some utterances are pragmatically more plausible than others.[1]

This paper describes machine learning experiments that combine acoustic confidences with move evaluation scores to classify n-best recognition hypothesis of spoken chess move instructions (e.g. “pawn e2 e4”) as correctly or incorrectly recognised. Classifying the n-best recognition hypotheses instead of the single-best (e.g. (Walker et al., 2000)) has the advantage that a correct hypothesis can be accepted even when it is not the highest scoring recognition result. Previous work on n-best hypothesis reordering (e.g. (Chotimongkol and Rudnicky, 2001)) has focused on selecting hypotheses with the lowest relative word error rate. In contrast, our approach makes predictions about whether hypotheses should be accepted or rejected. The learning experiments show significantly improved classification results over competitive baselines and underline the usefulness of incorporating higher-level information for utterances classification.

2 Domain and Data Collection

The domain of our research are spoken chess move instructions. We chose this scenario as a testbed for our approach for three main reasons. First, we can use move evaluation scores computed by a computer chess program as a measure of the pragmatic plausibility of hypotheses. Second, the domain is simple and allows us to collect data in a controlled way (e.g. we can control for player strength), and third, the domain is restricted in the sense that there is only a finite set of possible legal moves in every situation. Similar considerations already let researchers in the 1970s choose chess-playing as an example scenario for the HEARSAY integrated speech understanding system (Reddy and Newell, 1974).

We collected spoken chess move instructions in a small experiment from six pairs of chess players. All subjects were German native speakers and familiar with the rules of chess. The subject’s task was to re-play chess games (given to them as graphical representations) by instructing each other to move pieces on the board. Altogether, we collected 1978 move instructions under different experimental conditions (e.g. strong games vs. weak games) in the following four data sets: 1) language model, 2) training, 3) development, and 4) test.

The recordings of the language model games were transcribed and served to construct a context free recognition grammar for the Nuance 8.0[2] speech recogniser which was then used to process all other move instructions with 10-best output.

[1] Although it is possible to use dialogue-state dependent recognition grammars that reflect expectations of what the user is likely to say next, these expectations do not say anything about the plausibility of hypotheses.

[2] http://www.nuance.com. We thank Nuance Inc. for making their speech recognition software available to us.

3 Baseline Systems

The general aim of our experiments is to decide whether a recognised move instruction is the one intended by the speaker. A system should accept correct recognition hypotheses and reject incorrect ones. We define the following two baseline systems for this binary decision task.

3.1 First Hypothesis Baseline

The first hypothesis baseline uses a confidence rejection threshold to decide whether the best recognition hypothesis should be accepted or rejected. To find an optimal value, we linearly vary the confidence threshold returned by the Nuance 8.0 recogniser (integral values in the range $[0, 100]$) and use it to classify the training and development data.

The best performing confidence threshold on the combined training and development data was 17 with an accuracy of 63.8%. This low confidence threshold turned out to be equal to the majority class baseline which is to classify all hypotheses as correctly recognised. In order to get a more balanced distribution of classification errors, we also optimised the confidence threshold according to the cost measure defined in Section 5. According to this measure, the optimal confidence rejection threshold is 45 with a classification accuracy of 60.5%.[3]

3.2 First Legal Move Baseline

The first legal move baseline makes use of the constraint that user utterances only contain moves that are legal in the current board configuration. We thus first eliminate all hypotheses that denote illegal moves from the 10-best output and then apply a confidence rejection threshold to decide whether the best legal hypothesis should be accepted or rejected.

The best performing confidence threshold on the combined training and test data for the first legal move baseline was 23 with an accuracy of 92.4%. This threshold also optimised the cost measure defined in Section 5. The performance of both baseline systems on the test data is reported below in Table 2 together with the results for the machine learning experiments.

4 ML Experiments

We devise two different machine learning experiments for selecting hypotheses from the recogniser's n-best output and from a list of all legal moves given a certain board configuration.

In Experiment 1, we first filter out all illegal moves from the n-best recognition results and represent the remaining legal moves in terms of 32 dimensional feature vectors including acoustic confidence scores from the recogniser as well as move evaluation scores from a computer chess program. We then use machine learners to decide for each move hypothesis whether it was the one intended by the speaker. If more than one hypothesis is classified as correct, we pick the one with the highest acoustic confidence. If there is no legal move among the recognition hypotheses or all hypotheses are classified as incorrect, the input is rejected.

Experiment 2 adds a second classification step to Experiment 1. In case an utterance is rejected in Experiment 1, we try to find the intended move among all legal moves in the current situation. This is again defined in terms of a classification problem. All legal moves are represented by 31 dimensional feature vectors that include “similarity features” with respect to the interpretation of the best recognition hypothesis and move evaluation scores. Each move is then classified as either correct or incorrect. We pick a move if it is the only one that is classified as correct and all others as incorrect; otherwise the input is rejected. The average number of legal moves in the development and training games was 35.3 with a maximum of 61.

4.1 Feature Sets

The feature set for the classification of legal move hypotheses in the recogniser's n-best list (Experiment 1) consists of 32 features that can be coarsely grouped into six categories (see below). All features were automatically extracted or computed from the output of the speech recogniser, move evaluation scores, and game logs.

1. Recognition statistics (3): position in n-best list; relative position among and total number of legal moves in n-best list
2. Acoustic confidences (6): overall acoustic confidence; min, max, mean, variance, standard deviation of individual word confidences
3. Text (1): hypothesis length (in words)
4. Depth1 plausibility (10): raw & normalised move evaluation score wrt. scores for all legal moves; score rank; raw score difference to max score; min, max, mean of raw scores; raw z-score; move evaluation rank & z-score among n-best legal moves
5. Depth10 plausibility (10): same features as for depth1 plausibility (at search depth 10)
6. Game (2): ELO (strength) of player; ply number

The feature set for the classification of all legal moves in Experiment 2 is summarised below. Each move is represented in terms of 31 (automatically derived) features which can again be grouped into 6 different categories.

1. Similarity (5): difference size; difference bags; overlap size; overlap bag
2. Acoustic confidences (6): same as in Experiment 1 for best recognition hypothesis

[3] 45 is also the default confidence rejection threshold of the Nuance 8.0 speech recogniser.

3. Text (2): length of best recognition hypothesis (in words) and recognised string (bag of words)
4. Depth1 plausibility (8): same as in Experiment 1 (w/o features relating to n-best legal moves)
5. Depth10 plausibility (8): same as in Experiment 1 (w/o features relating to n-best legal moves)
6. Game (2): same as in Experiment 1

The similarity features are meant to represent how close a move is to the interpretation of the best recognition result. The motivation for these features is that the machine learner might find regularities about what likely confusions arise in the data. For example, the letters "b", "c", "d", "e" and "g" are phonemically similar in German (as are the letters "a" and "h" and the two digits "zwei" and "drei"). Although the move representations are abstractions from the actual verbalisations, the language model data showed that most of the subjects referred to coordinates with single letters and digits and therefore there is some correspondence between the abstract representations and what was actually said.

4.2 Learners

We considered three different machine learners for the two classification tasks: the memory-based learner TiMBL (Daelemans et al., 2002), the rule induction learner RIPPER (Cohen, 1995), and an implementation of Support Vector Machines, SVM^{light} (Joachims, 1999). We trained all learners with various parameter settings on the training data and tested them on the development data. The best results for the first task (selecting legal moves from n-best lists) were achieved with SVM^{light} whereas RIPPER outperformed the other two learners on the second task (selecting from all possible legal moves). SVM^{light} and RIPPER where therefore chosen to classify the test data in the actual experiments.

5 Results and Evaluation

5.1 Cost Measure

We evaluate the task of selecting correct hypotheses with two different metrics: i) classification accuracy and ii) a simple cost measure that computes a score for different classifications on the basis of their confusion matrices. Table 1 shows how we derived costs from the *additional* number of steps (verbal and non-verbal) that have to be taken in order to carry out a user move instruction. Note that the cost measure is not validated against user judgements and should therefore only be considered an indicator for the (relative) quality of a classification.

5.2 Results

Table 2 reports the raw classification results for the different baselines and machine learning experiments together with their accuracy and associated cost. Here and in subsequent tables, FH$_{17}$ and FH$_{45}$ refer to the first hypothesis baselines with confidence thresholds 17 and 45 respectively, FLM to the first legal move baseline, and Exp1 and Exp2 to Experiments 1 and 2 respectively.

Class	**Cost**	**Sequence**
accept correct	0	instruct – move
reject correct/ reject incorrect	2	instruct – *reject* – *instruct* – move
accept incorrect	4	instruct – *move* – *object* – *move* – *instruct* – move

Table 1: Cost measure

	accept	**reject**
	FH$_{17}$ (Acc: 61.7% Cost: 1230)	
correct	489	0
incorrect	306	3
	FH$_{45}$ (Acc: 64.3% Cost: 1188)	
correct	441	48
incorrect	237	72
	FLM (Acc: 93.5% Cost: 358)	
correct	671	0
incorrect	52	75
	Exp1 (Acc: 97.2% Cost: 246)	
correct	695	2
incorrect	20	81
	Exp2 (Acc: 97.2% Cost: 176)	
correct	731	1
incorrect	21	45

Table 2: Raw classification results

The most striking result in Table 2 is the huge classification improvement between the first hypothesis and the first legal move baselines. For our domain, this shows a clear advantage of n-best recognition processing filtered with "hard" domain constraints (i.e. legal moves) over single-best processing.

Note that the results for Exp1 and Exp2 in Table 2 are given "by utterance" (i.e. they do not reflect the classification performance for individual hypotheses from the n-best lists and the lists of all legal moves). Note also that both the different baselines and the machine learning systems have access to different information sources and therefore what counts as correctly or incorrectly classified varies. For example, the gold standard for the first hypothesis baseline only considers the best recognition result for each move instruction. If this is not the one intended by the speaker, it counts as *incorrect* in the gold standard. On the other hand, the first legal move among the 10-best recognition hypotheses for the same utterance might well be the correct one and would therefore count as *correct* in the gold standard for the FLM baseline.

5.3 Comparing Classification Systems

We use the χ^2 test of independence to compute whether the classification results are significantly different from each other. Table 3 reports significance results for comparing the different classifications of the test data. The table entries include the differences in cost and the level of statistical difference between the confusion matrices as computed by the χ^2 statistics (*** denotes significance at $p = .001$, ** at $p = .01$, and * at $p = .05$). The table should be read row by row. For example, the top row in Table 3 compares the classification from Exp2 to all other classifications. The value -1054^{***} means that the cost compared to FH_{17} is reduced by 1054 and that the confusion matrices are significantly different at $p = .001$.

	$\mathbf{FH_{17}}$	$\mathbf{FH_{45}}$	**FLM**	**Exp1**
Exp2	-1054^{***}	-1012^{***}	-182^{***}	-70^{**}
Exp1	-984^{***}	-942^{***}	-112^{***}	
FLM	-872^{***}	-830^{***}		
$\mathbf{FH_{45}}$	-42^{***}			

Table 3: Cost comparisons and χ^2 levels of significance for all test games

Tables 4 and 5 compare the performance of the different systems for strong and weak games (a variable controlled for during data collection).

	$\mathbf{FH_{17}}$	$\mathbf{FH_{45}}$	**FLM**	**Exp1**
Exp2	-568^{***}	-568^{***}	-102^{***}	-40^{*}
Exp1	-528^{***}	-528^{***}	-62^{**}	
FLM	-466^{***}	-466^{***}		
$\mathbf{FH_{45}}$	$\pm 0^{***}$			

Table 4: Cost comparisons and χ^2 levels of significance for strong test games

	$\mathbf{FH_{17}}$	$\mathbf{FH_{45}}$	**FLM**	**Exp1**
Exp2	-486^{***}	-444^{***}	-80^{*}	-30
Exp1	-456^{***}	-414^{***}	-50	
FLM	-406^{***}	-364^{***}		
$\mathbf{FH_{45}}$	-42^{***}			

Table 5: Cost comparisons and χ^2 levels of significance for weak test games

The results show that the machine learning systems perform better for the strong test data. We conjecture that the poorer results for the weak data are due to more bad moves in these games which receive a low evaluation score and might therefore be considered incorrect by the learners.

6 Conclusions

We presented a machine learning approach that combines acoustic confidence scores with automatic move evaluations for selecting from the n-best speech recognition hypotheses in a chess playing scenario and compared the results to two different baselines.

The chess scenario is well suited for our experiments because it allowed us to filter out impossible moves and to use a computer chess program to assess the plausibility of legal moves. However, the methodology underlying Experiment 1 can be applied to other spoken dialogue systems to choose interpretation(s) from a recogniser's n-best output. We have successfully used this setup for classifying hypotheses in a command and control spoken dialogue system (Gabsdil and Lemon, subm). Experiment 2 exploits the fact that the number of possible interpretations is finite in the chess scenario. Although this obviously does not hold for many dialogue tasks, there are applications such as call routing (e.g. (Walker et al., 2000)) where the number of possible interpretations is limited in a similar way. Instead of selecting correct interpretations, we imagine that one could also use the proposed setup to decide which of a finite set of dialogue moves was performed by a speaker.

References

Ananlada Chotimongkol and Alexander I. Rudnicky. 2001. N-best Speech Hypotheses Reordering Using Linear Regression. In *Proceedings of EuroSpeech-01*.

William W. Cohen. 1995. Fast Effective Rule Induction. In *Proceedings of ICML-95*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2002. TIMBL: Tilburg Memory Based Learner, version 4.2, Reference Guide. Available from `http://ilk.kub.nl/downloads/pub/papers/ilk0201.ps.gz`.

Malte Gabsdil and Oliver Lemon. subm. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. Submitted to ACL-04.

Thorsten Joachims. 1999. Making Large-Scale SVM Learning Practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 41–55. MIT Press.

R. Reddy and A. Newell. 1974. Knowledge and its representation in a speech understanding system. In L.W. Gregg, editor, *Knowledge and Cognition*.

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-00*.