A RERANKING APPROACH FOR RECOGNITION AND CLASSIFICATION OF SPEECH INPUT IN CONVERSATIONAL DIALOGUE SYSTEMS

Fabrizio Morbini *Kartik Audhkhasi* *Ron Artstein* *Maarten Van Segbroeck* *Kenji Sagae*
Panayiotis Georgiou *David R. Traum* *Shri Narayanan*

University of Southern California
Los Angeles, CA 90094, USA
{morbini,artstein,sagae,traum}@ict.usc.edu
{audhkhas,vansegbr}@usc.edu, {georgiou,shri}@sipi.usc.edu

ABSTRACT

We address the challenge of interpreting spoken input in a conversational dialogue system with an approach that aims to exploit the close relationship between the tasks of speech recognition and language understanding through joint modeling of these two tasks. Instead of using a standard pipeline approach where the output of a speech recognizer is the input of a language understanding module, we merge multiple speech recognition and utterance classification hypotheses into one list to be processed by a joint reranking model. We obtain substantially improved performance in language understanding in experiments with thousands of user utterances collected from a deployed spoken dialogue system.

Index Terms— Automatic speech recognition, Interactive systems, Natural language processing, Supervised learning

1. INTRODUCTION

Many spoken dialogue systems are based on a pipeline structure that includes separate modules for automatic speech recognition (ASR), natural language understanding (NLU), dialogue management, language generation and speech synthesis. Once speech input is received, the output of each of these modules is passed as input to the next module until speech output is produced. Although more sophisticated alternatives exist, e.g. to achieve incremental processing [1] or to support sophisticated cognitive architectures [2], the simple pipeline view between ASR and NLU is widely used, and is assumed in popular spoken dialogue system toolkits, such as the CSLU Toolkit [3] and the CMU Olympus Toolkit [4].

The pipeline assumption between ASR and NLU allows for greater modularity in the system, with ASR being responsible exclusively for transcribing the audio signal as text, and NLU being responsible exclusively for interpreting text input. However, perfect transcription of spoken user input is beyond the capabilities of current ASR technology, especially in conversational systems, such as virtual human dialogue systems [5, 2]. In many virtual human systems, speech recognition errors can be tied directly to degradation in system performance [6].

Motivated by the idea that recognition of the words in an utterance and interpretation of that utterance are two processes that are more intimately connected than a simple unidirectional pipeline suggests, we examine the use of combined information from ASR and NLU in a reranking framework that attempts to model aspects of the two tasks jointly. More specifically, we rescore speech hypotheses in the form of n-best lists generated by one or more ASR engines by taking the NLU interpretation of these hypotheses into account. In contrast to approaches that aim to improve speech recognition, for example by correcting ASR output [7], reranking ASR n-best lists [8] or rescoring lattices [9], our approach is focused on improving the dialogue system's capability to understand speech input, not word error rates (WER). In the experiments presented in this paper, we assume a simple form of natural language understanding for dialogue systems that consists of assigning a category label to each user input utterance. The category labels in the dialogue system correspond roughly to dialogue acts. However, the overall approach we present is flexible with respect to NLU approach and representation, and could be applied in a variety of situations involving different systems with spoken input.

We note that the work described here is not intended as a method to improve word error rate in automatic speech recognition; our goal, instead, is to improve the accuracy of natural language understanding in spoken dialogue systems. The main contributions of this paper are: **1)** we show that using NLU information to rerank ASR n-best lists brings a significant increase in NLU performance compared to a pipeline approach, even when that pipeline includes a discriminative language model based on a reranker without NLU information; and **2)** reranking can take advantage of multiple n-best lists obtained from different ASR engines.

This paper is organized as follows: in section 2 we de-

scribe the data used to develop and evaluate our approach. In section 3 we describe our experiments and results. Finally we conclude in section 4.

2. DATA

We use a subset of the Twins corpus [10] that contains speech files that are transcribed and manually annotated with correct interpretations in the context of a virtual human dialogue system that implements two virtual museum guides. Each audio file, containing a question asked by a museum visitor [11], was transcribed and annotated with the set of appropriate system responses. Because this dialogue system implements a simple reflex agent that produces a response based only on a classification of the input utterance, the annotation of appropriate system responses can be viewed as the interpretation of the input utterances. In this system, the natural language understanding module classifies the input utterance into categories. Typically these categories correspond to dialogue acts with domain-specific semantics, e.g. [12], or sometimes to ad-hoc equivalence classes designed to map arbitrary natural language input to a closed set of system actions [5]. Our annotated subset of the Twins corpus includes audio files, gold-standard text transcriptions and manually annotated NLU category labels for each utterance.

The data consists of 13,908 audio files, corresponding to 2,746 unique transcribed utterances representing a total of 168 unique interpretation labels, each corresponding to a specific type of question or statement that museum visitors might say to the virtual museum guides, except for one category label, which is used to categorize utterances that are outside of the domain covered by the system (we refer to this category as *off-topic*). The distribution of utterances is highly skewed (see figure 1), with almost a quarter of user utterances representing greetings. The dataset is partitioned into training, development and testing sets, with utterances collected on a single day always assigned to the same partition; about 70% of the utterances are in the training set and the remaining ones are equally split into the development and testing sets.

2.1. Overview of data preparation

To achieve our goal of developing and testing techniques to improve natural langauge understanding in spoken dialogue systems by joint discriminative reranking of speech n-best lists and natural language understanding, we start by processing audio utterances in our dataset using three ASR engines to produce three n-best lists per utterance. We then automatically annotate each hypothesis in each speech n-best list with the corresponding NLU output. The result is a dataset containing, for each utterance, multiple ASR n-best lists, with each ASR hypothesis annotated with the 1-best NLU label. The following sections describe this process in more detail.

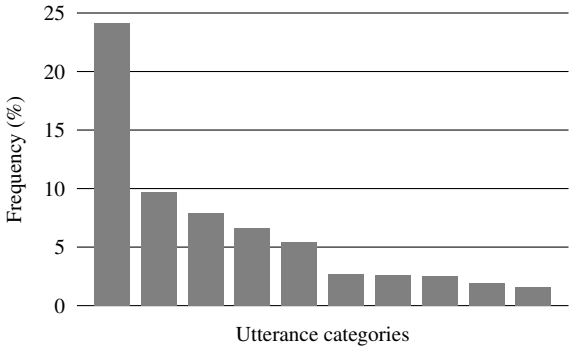

Fig. 1. Distribution of the 10 most common NLU labels in the Twins data set. Almost a quarter of user utterances are greetings.

2.2. Data preparation: speech n-best lists

For the experiments reported in this paper, we used three state of the art and readily available ASR engines: AT&T's Watson, available through the AT&T Speech Mashup service, the Google Speech API, and SAIL's[1] OtoSense-Kaldi. Watson, the AT&T recognizer, requires that a custom language model (LM) be built from a text file, while the Google ASR does not allow for any customization. The Google ASR also limits the maximum file length to roughly 10 seconds of audio. This did not significantly affect our results, since only a small fraction of the utterances in our dataset have durations greater than 10 seconds.

The recently developed SAIL real-time ASR engine, OtoSense-Kaldi, is built on top of the open-source toolkit, Kaldi [13]. Context-dependent tri-phone acoustic models were trained on 39-dimensional per-frame feature vectors composed of Mel Frequency Cepstral Coefficients (MFCCs) with delta and double-delta features. The total number of tied Gaussians was set to 10000. After training the context-dependent models, we used Linear Discriminant Analysis (LDA) and 4 iterations of Maximum Mutual Information (MMI) training for discriminatively updating the acoustic model parameters. MMI training updates the parameters by maximizing the mutual information between the observation sequence and the correct state sequence. We observed that discriminative training gives an improvement of approximately 6% (relative) in WER on the held out test set, resulting in a final WER on the test set of 20.9%. Note for reference, that WSJ models for this dataset (with the same LM) result in an error rate of around 50%, which indicates the large mismatch in the acoustic conditions and speaking style. We believe that additional gains can be made through employing other acoustic data corpora beyond this dataset, model selection and on-line adaptation.

The language models employed by OtoSense-Kaldi and

[1] http://sail.usc.edu

ASR engine	1-best WER	Oracle WER
AT&T	28.8	16.8
Google	24.4	17.2
OtoSense-Kaldi	18.7	11.5

Table 1. Word error rates obtained by processing the development set with the three ASR engines. These WERs are obtained dividing the total number of edits by the total number of words in the transcriptions.

AT&T[2] engines were the same. The LM of the Google system was beyond our control. However, we did notice that during subsequent decodings of the test set the WER went from 30.6% down to 24.4%. This may be due to internal data mining and refining by Google, that may indicate some bias in the results.

On the acoustic modeling front, Otosense-Kaldi was trained using only this data. For the purposes of this publication, we did not train any gender or age specific models. AT&T provides a limited set of acoustic models, we used the default `gentel06` model, while Google's acoustic models are again automatically selected.

We ran the training, development and testing sets through the ASR systems and recorded for each utterance the returned n-best lists (with $n = 30$). To generate the custom language model for the AT&T and OtoSense-Kaldi ASR we used the following procedure: for the **training** set, we divided it into 10 folds and for each fold, we used the manual transcriptions found in the other 9 folds to generate the text file used to train the language model; for the **development** and **testing** sets we simply used the manual transcriptions in the training set to generate the files used to build the language models.

Table 1 shows the word error rates obtained in our development set by the three ASR systems used. All perform well, with OtoSense-Kaldi achieving the highest performance. However, for the purpose of this paper, the performance of the ASR engines used is not crucial as our primary focus in on NLU accuracy, as impacted by reranking n-best lists from multiple ASR engines jointly with NLU information.

2.3. Data preparation: utterance classification

As a first-pass natural language understanding module to categorize each of the speech hypotheses contained in the n-best lists produced from the speech data, we used a maximum entropy multiclass classifier with lexical features (unigrams and bigrams) [14]. To label the speech n-best lists in our training set with this NLU classifier, we performed a 10-fold cross-validation: for each fold, we trained the maximum entropy classifier using the transcriptions and the manually annotated NLU categories in the other nine folds. To label the speech n-best lists in our development and test sets with the NLU classifier, we trained the NLU using the text and NLU categories in the entire training set.

[2]While we do not have internal knowledge of how the AT&T system works, we supplied both systems with the same data. We are unsure of any other resources AT&T may use for training the LM.

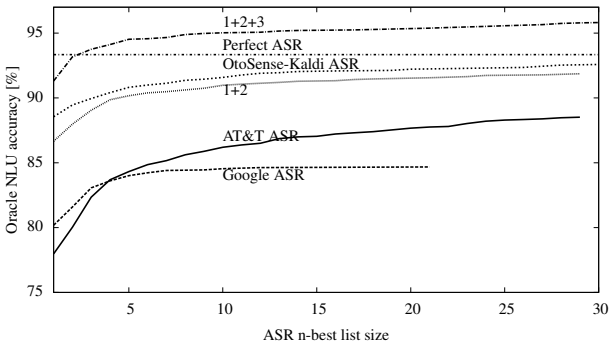

Fig. 2. The baseline NLU performance on the development set achieved using n-best lists obtained from the three ASRs. The curves show the maximum gains made possible by considering longer n-best lists. Perfect ASR uses manual transcriptions for training and testing the NLU. All other curves test the NLU on ASR output. *1+2* shows the NLU performance after combining the n-best lists from AT&T and Google ASRs. *1+2+3* shows the NLU performance after combining all three ASRs.

After these steps our dataset is composed of a set of audio files, each annotated with **1)** a manual transcription; **2)** a manually assigned gold-standard NLU label and **3)** three speech n-best lists with each speech hypothesis annotated with the 1-best result of the automatic NLU classification. Items 1) and 2) are used for training models and as reference labels for evaluation. The information in item 3) is produced fully automatically from the audio files. It is this information in the development and test sets, which is composed from the results of ASR and NLU trained on the training set, that we attempt to rescore to improve overall NLU performance.

3. EXPERIMENTS

We evaluate different ways to improve on the simple pipeline where a single speech hypothesis is produced by the speech recognizer and used as input to the natural language understanding module. In our experiments, we use the data described above, and the main task under consideration is the classification of individual utterances as one out of the 168 interpretation categories (including the *off-topic* category) available in the museum guides system. NLU performance is evaluated using an accuracy metric, calculated as the number of correct category assignments divided by the number of utterances in the test data.

Figure 2 shows NLU accuracy on the development data

under several conditions. The *Perfect ASR* condition corresponds to what NLU accuracy would be if there were no ASR errors[3]. This estimate is obtained using the manual transcriptions in the corpus as the input to the NLU module. The other five curves show the upper-bound on NLU accuracy for different sizes of n-best lists and different ASR engines and their combinations, by assuming an oracle that can pick the best speech hypothesis out of the n-best list (i.e. the hypothesis that results in highest NLU accuracy, which may not be the hypothesis with lowest word error rate). The *1+2* curve uses for each audio file a n-best list formed by merging the two n-best lists obtained from the AT&T and Google ASR engines. This merged n-best list is built by interleaving the two original n-best lists. Similarly, the *1+2+3* curve gives for the result obtained by merging the three n-best lists[4].

This figure shows that using speech n-best lists has the potential for considerable improvements over 1-best ASR output, but that even better results could be obtained if hypotheses generated by different recognizers are combined, if we had a way to approximate the oracles mentioned above.

3.1. Classification of reranked speech results

Figure 2 shows that picking the right speech hypothesis from an ASR n-best list can result in large improvements in NLU accuracy. We first address this task using a discriminative language model trained using the perceptron algorithm with unigram, bigram and trigram features, following [15, 8]. In tables 2 and 3 this method is listed as *WER reranking*. In almost all cases this method achieved an improved NLU accuracy even though in few cases the WER actually became worse. The highest improvement in both NLU accuracy and WER is achieved when combining the n-best lists from all ASRs, although this method cannot reach the 1-best performance achieved by the best performing ASR (OtoSense-Kaldi) when considered alone.

3.2. Joint rescoring of speech and utterance classification

In this section, we explore the use of feature sets that encode information obtained from all ASR hypotheses and their NLU labels, according to the first-pass NLU classifier.

Given an ASR n-best list where each speech hypothesis has been annotated automatically with an NLU label, we derive a new k-best list, where k is the number of different NLU labels in the initial ASR n-best list. We then use the same reranking approach as in the previous section, training with a 0-1 loss function that reflects whether or not the i^{th} entry in the k-best list corresponds to the gold-standard reference NLU label. After experimenting with several feature sets using our training and development data, we arrived at the following features to rescore an entry i in this k-best list:

NGrams: the unigrams, bigrams and trigrams from all of the ASR hypotheses in the original speech n-best list annotated with the same NLU label l_i;

NLU: the NLU label l_i;

RelPos: the position in the speech n-best list of the topmost ASR result that was labeled with l_i by the first-pass NLU (10 bins);

Pos: six binary features that indicate whether the topmost ASR hypothesis in the speech n-best list annotated with l_i has the rank of first, second, third, fourth, fifth or lower in the speech n-best list;

Count: the frequency with which the label l_i appeared in the ASR n-best list (10 bins);

OffTopic: a binary feature that reflects whether l_i is the *off-topic* label.

Applying this NLU rescoring approach to the n-best list obtained from one ASR engine for each utterance in the development set results in a higher improvement in NLU accuracy when compared to the WER reranker. the improvement increases when we consider n-best lists formed by merging results from several ASR engines. This joint reranker is able to improve over the 1-best performance achieved by OtoSense-Kaldi in the development set, however in the testing set the improvement is not significant.

Table 3 shows the performance of the approaches described here on the test set, compared to their respective 1-best performance. All NLU accuracy differences greater than 1% are significant with $p < 0.03$[5].

3.3. Contribution of individual features

Here we describe our findings about the contribution to the NLU accuracy of each feature used by the NLU reranker described in section 3.2.

Table 4 shows the importance of the individual features used when compared with the baseline system[6]. NGram features are always present. The most valuable feature is the Count feature. When added to the NGram features it gives an increment in performance (average across all runs 2.8%) that is significant in 90% of runs (one run correspond to the evaluation of the significance of the performance difference

[3]Perfect ASR is independent of the n-best list size as it uses only the transcriptions, but in this figure is displayed as a line to facilitate comparisons.

[4]The *1+2+3* curve achieves a oracle NLU performance higher than the perfect ASR case. This can be explained by considering that this longer and more diverse combined n-best list can contain more different NLU labels, therefore increasing the chance of finding a match with the gold NLU label.

[5]Using Dan Bikel's stratified shuffling significance test script: http://www.cis.upenn.edu/~dbikel/software.html#comparator

[6]We based this analysis, for simplicity, on a subset of the available cases: the development set annotated with n-best lists from the AT&T and Google ASRs.

Method	NLU accuracy [%]	
Perfect ASR	93.3	
		Δ WER
WER reranking, AT&T	78 (+2.1)	-0.6
WER reranking, Google	79.8 (+1)	-0.2
WER reranking, OtoSense	85.6 (-0.6)	+1.7
WER reranking, 1+2	83.3 (+7.4/+4.5)	-7.1/-2.7
WER reranking, 1+2+3	85 (+9.1/+6.2/-1.2)	-7.5/-3.1/+2.6
		N-best
NLU reranking, AT&T	79.7 (+3.8)	16
NLU reranking, Google	80.1 (+1.3)	5
NLU reranking, OtoSense	86.2 (0)	1
NLU reranking, 1+2	84.4 (+8.5/+5.6)	7
NLU reranking, 1+2+3	88.1 (+12.2/+9.3/+1.9)	6

Table 2. Performance of the reranking techniques on the development set. The second column shows the NLU accuracy achieved and in brackets we report the difference with the NLU accuracy obtained by using the correspondent 1-best ASR result (for the combined case we report the difference with the individual ASRs). The third column shows for the WER reranking results the change in WER with respect to the correspondent 1-best WER. For the NLU reranking we report the n-best list size that when re-ranked achieved the best NLU accuracy (reported in column 2). All NLU accuracy differences greater than 1% are significant with $p < 0.03$.

Method	NLU accuracy [%]	
Perfect ASR	91.2	
		Δ WER
WER reranking, AT&T	71.9 (+1.9)	-0.2
WER reranking, Google	76.2 (+1.6)	+0.4
WER reranking, OtoSense	81 (-1)	+2.6
WER reranking, 1+2	78.4 (+8.4/+3.8)	-8.6/-0.8
WER reranking, 1+2+3	80 (+10/+5.4/-2)	-7.1/+.7/+4.7
		N-best
NLU reranking, AT&T	73.3 (+3.3)	16
NLU reranking, Google	76.6 (+2)	5
NLU reranking, OtoSense	81.8 (-0.2)	1
NLU reranking, 1+2	79.5 (+9.5/+4.9)	7
NLU reranking, 1+2+3	82.6 (+12.6/+8/+0.6)	6

Table 3. Performance of the rescoring techniques described on the testing set. We run the NLU rescoring using the N-best list sizes that gave the best performance on the development set (see table 2). All NLU accuracy differences greater than 1% are significant with $p < 0.03$.

between the baseline and the reranker for a particular n-best length).

The least valuable feature is the one based on the off topic label.

Overall the NLU features are less effective when considered by themselves than the features purely based on the ASR n-best list. However, the combination of both types of features (NLU based and ASR based) achieves higher performance than each group individually.

Feature	Average significance [%]	Average Δ NLU accuracy [%]
Count	90	2.8
Pos	89.6	2.0
RelPos	80.8	3.2
NLU	68.7	2.1
OffTopic	41.2	1.6

Table 4. For each row, significance was measured when $p <= 5\%$ between the performance of the baseline and that obtained by the reranker using the NGram feature plus the feature associated to that row.

3.4. Error analysis

In this section we go in more detail into the change in NLU performance obtained by the system that rescores based on ASR and NLU information. Also here for simplicity we limit ourselves to the development set annotated by the AT&T ASR. The rescoring model changed 161 interpretations of which 68% were correct changes. Most changes from correct to incorrect NLU analysis involved the *off-topic* label, with the rescoring model having more problems recognizing when an utterance should be considered off-topic.

Using the Google ASR n-best lists, the model changed 144 interpretations of which 78% were correct changes. Also in this case most errors happened with failing to recognize an utterance as off-topic.

For the combined *1+2* case, the model changed 222 interpretations of which 84% were correct changes. Here too the category that introduced most errors was the off-topic one.

Concluding, the rescoring model is effective in improving the NLU interpretation, but is biased toward avoiding the off-topic label. This is partly due to the off-topic category encompassing a very broad class of utterances for which collecting a comprehensive training set is practically impossible (unless the domain's topic strongly limits the typical off-topic utterances, and in that case those limited and highly probable off-topic utterances should be added to the domain).

4. CONCLUSION

We have shown that a rescoring framework that integrates information from both ASR and NLU, optimizing the use of ASR n-best lists directly for NLU accuracy, outperforms one that focuses on optimizing word error rate as part of a strict ASR-NLU pipeline. In addition, we have shown that rescoring is effective in utilizing n-best lists obtained from multiple ASR systems. In fact our rescoring approach achieves its highest improvement in NLU accuracy when reranking the n-best lists obtained by merging the results of all three recogniz-

ers. This is particularly interesting given that all ASR systems perform very well on their own, showing that our reranking approach is a promising framework for ASR ensembles where results from multiple recognizers are used effectively.

Next we plan to run experiments with more challenging speech signals to test whether our reranking technique works well also with higher word error rates. We would also like to investigate the use of this approach with more sophisticated NLU models, other ASR engines, and with different types of spoken dialogue systems.

5. REFERENCES

[1] David Schlangen and Gabriel Skantze, "A general, abstract model of incremental dialogue processing," *Dialogue & Discourse*, vol. 2, no. 1, pp. 83–111, 2011.

[2] David R. Traum, William Swartout, Jonathan Gratch, and Stacy Marsella, *A Virtual Human Dialogue Model for Non-team Interaction*, Springer, 2008.

[3] Stephen Sutton, Ronald Cole, Jacques De Villiers, Johan Schalkwyk, Pieter Vermeulen, Mike Macon, Yonghong Yan, Ed Kaiser, Brian Rundle, Khaldoun Shobaki, Paul Hosom, Alex Kain, Johan, Johan Wouters, Dominic Massaro, and Michael Cohen, "Universal Speech Tools: The Cslu Toolkit," in *In Proceedings of the International Conference on Spoken Language Processing (ICSLP*, 1998, vol. 7, pp. 3221–3224.

[4] Dan Bohus, Antoine Raux, Thomas K. Harris, Maxine Eskenazi, and Alexander I. Rudnicky, "Olympus: an open-source framework fro conversational spoken language interface research," in *proceedings of HLT-NAACL 2007 workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, 2007.

[5] Anton Leuski and David R. Traum, "Practical language processing for virtual humans," in *Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*, 2010.

[6] William Yang Wang, Ron Artstein, Anton Leuski, and David R. Traum, "Improving spoken dialogue understanding using phonetic mixture models.," in *FLAIRS Conference*, R. Charles Murray and Philip M. McCarthy, Eds. 2011, AAAI Press.

[7] Eric K. Ringger and James F. Allen, "Robust error correction of continuous speech recognition," in *In Proceedings of the ESCA-NATO Robust Workshop*, 1997.

[8] Brian Roark, Murat Saraclar, and Michael Collins, "Discriminative n-gram language modeling," *Computer Speech & Language*, vol. 21, no. 2, pp. 373 – 392, 2007.

[9] A. Rastrow, M. Dreyer, A. Sethy, S. Khudanpur, B. Ramabhadran, and M. Dredze, "Hill climbing on speech lattices: A new rescoring framework," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 5032 –5035.

[10] Priti Aggarwal, Ron Artstein, Jillian Gerten, Athanasios Katsamanis, Shrikanth Narayanan, Angela Nazarian, and David Traum, "The Twins corpus of museum visitor questions," in *LREC-2012*, Istanbul, Turkey, May 2012.

[11] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, Chad Lane, Jacquelyn Morie, Priti Aggarwal, Matt Liewer, Jen-Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White, "Ada and Grace: Toward realistic and engaging virtual museum guides," in *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20–22, 2010 Proceedings*, Jan Allbeck, Norman Badler, Timothy Bickmore, and Alla Pelachaud, Catherine Safonova, Eds., vol. 6356 of *Lecture Notes in Artificial Intelligence*, pp. 286–300. Springer, Heidelberg, 2010.

[12] Ron Artstein, Sudeep Gandhe, Michael Rushforth, and David R. Traum, "Viability of a simple dialogue act scheme for a tactical questioning dialogue system," in *DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue*, Stockholm, Sweden, June 2009, p. 43–50.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[14] Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum, "Towards natural language understanding of partial speech recognition results in dialogue systems," in *Short Paper Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 conference*, 2009.

[15] Brian Roark, Murat Saraclar, Michael Collins, and Mark Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of ACL*, 2004.