Improving Spoken Dialog Systems

Benjamin Han

Language Technologies Institute
Carnegie Mellon University

benhdj@cs.cmu.edu

December 14, 2000

Abstract

In this paper the possible sources of inaccuracies of a spoken dialog system are discussed, and the approaches to address each of them are surveyed.

A complete spoken dialog system consists of a speech recognizer, a linguistic processor (including a parser), a dialog manager and a text/speech generator. In the first three components inaccuracies might arise from recognition errors, insufficient or inaccurate linguistic knowledge (e.g. insufficient grammar coverage), and less than optimal dialog strategies deployed. In this paper each of the sources of the inaccuracies is discussed and various techniques to address them are surveyed.

1 Recognition Errors

Although it is possible to further tune a recognizer and re-train the language model to enhance the effectiveness of a recognizer, the working assumption has been to treat it as a black box, thus only the end result of the recognition is available for correction. Most of the works discussed assumed the output of the recognizer to be an N-best list of utterance hypotheses, and the correction involves re-ordering (re-scoring) them according to different knowledge sources [15, 10, 16]. Two papers reported selecting the best hypothesis directly from a word lattice. In [9] different paths in a lattice are first aligned, and a new utterance hypothesis is formed by observing the consensus between different paths. The authors reported better WER improvement over their previous N-best re-scoring approach. In [11] an arc in a lattice is dynamically scored based on a trained reliability model, which indicates how reliable an acoustic arc can be distinguished from its immediate competing arc.

It is worth noting that incorporating linguistic knowledge, although intuitively the right thing to do, gives only marginal WER improvement over the statistical approaches. One of the best results was reported in [15]: the WER on the sentences of unlimited length was reduced 2.5% absolute. Their work proposed a general framework for fusing multiple knowledge sources by using *discrimination functions* - functions that indicate how reliable a particular feature is able to distinguish between good and bad hypotheses. The authors gave a very interesting point in explaining why the approach worked: the system attempts directly to model characteristic mistakes made by the recognizer.

The work in [1] investigated the bottom-line question about how useful linguistic knowledge is to improve the recognition. Human subjects were given 10-best lists from a recognizer and asked to take advantage of a battery of linguistic tests to pick the most likely hypothesis. The reductions of WER were then credited to individual linguistic test, and an overall comparisons were shown between the oracle error rate (always pick the most correct one out of the 10-best lists), the human selection error rate and the human editing error rate (human subjects are allowed to edit the hypotheses as they see fit), among them the oracle error rate serves as an upper bound for any post-editing effort. The experiment was an attempt in the right direction to single out the linguistic tests with maximal expected utility, although some tests appeared to have inter-dependency (e.g., open class word choice vs. world knowledge) and thus need to be refined.

A related research topic is *confidence annotation*, which assigns an indicator to each word (word-level) or utterance (utterance-level) to indicate if the recognition is correct [2, 18]. Although this does not involve any correction effort after the annotation, it can be done if a richer representation of the indicators is given.

In summary it seems to be a reasonable conclusion that any useful post-editing effort should incorporate different knowledge sources at different levels: both at acoustic level, word level and sentence level. A uniform framework for fusing different knowledge sources is thus necessary for a more thorough theoretical and empirical analysis.

2 Insufficient or Inaccurate Linguistic Knowledge

Within a dialog system a linguistic processor is used for mapping the surface form of an utterance to a meaning representation. Optionally it can also be used to improve the recognition result by scoring a hypothesis based on the analysis (e.g., grammaticality, semantic coherence and completeness, etc.). For both tasks insufficient or inaccurate knowledge such as incomplete grammar coverage could lead to incorrect interpretation of the user intention, more dialog repair turns, or even a complete system breakdown. In spoken dialog systems the problems are real since speech tends to be fragmented and ungrammatical.

To remedy the problem a robust parser is needed to at least give a set of parsable chunks, which are then interpreted to give the final syntactic or semantic analyses. One of this kind of robust parsers is SOUP [7], which utilizes the probabilistic context-free grammar formalism and provides many convenient parsing modes for robust interpretation (e.g., skipping and all-top parsing mode). The parser serves as the backbone of the grammar learning system GSG [5, 6], which acquires new grammar rules from the interactions with the end user. GSG is based on a semantic grammar formalism, where the surface form is directly mapped into the corresponding meaning representation. The system first predicts the top-most node acting as the speech act of the utterance, based on the chunks it received from SOUP. The complete semantic parse tree is then constructed by observing several heuristics, including following the links in an ontology compiled from the grammar. During the process the interactions are initiated both by the system and by the user for clarification. However, the system lacks the ability to learn new concepts (non-terminals in the semantic grammar) from the interactions.

A different approach for robust parsing was reported in [17]. Based on genetic programming, parser ROSE evolves a population of parse trees by combining chunks using a single function MY-COMB. The optimal sequence of MY-COMB applications, yielding an optimal parse tree, is thus viewed as an evolved program. The most challenging problem encountered was training the fitness function for the selection process, because the optimal tree is obviously unknown a priori.

In addition to the problem of grammar coverage, insufficient lexical knowledge also poses a challenge for any linguistic analysis. In [20, 19] the authors reported using an information theoretic network to learn the meaning of words in a multi-sensory environment. Essentially the learning agent associates the visual input (camera) with the auditory input (microphone) to learn the meaning such as 'red'. In fact the association can be between any representation, thus the approach might be applicable to learning lexical semantics grounding in some meaning representation. However, one of the limitations, as admitted by the authors, was that the network formalism could not capture the more complex linguistic structures, therefore learning grammars could be a challenging task.

Another interesting issue is the integration of the linguistic processor and the recognizer. An obvious approach is to view each as a separate process: the linguistic processor takes the recognition result as an input, and outputs an analysis. Another approach is to integrate them at an earlier stage of the process. In [8] the GLR parser is extended to parse directly on a word lattice, and is capable of merging two partial analyses. Some works discussed in the previous section also use a linguistic processor closely with a recognizer in order to evaluate the recognition results [15, 10].

In summary an ideal linguistic processor in a spoken dialog system should be both robust and adaptive. To be adaptive, the processor needs to learn new rules either from passive observations, or by actively involving in the dialog management process in order to pose appropriate questions to the end user. Another interesting line of research is to investigate the interaction between ontologies and languages. The questions: how does an ontology assist language understanding, and how to build an ontology via interactions with a user?

3 Non-optimal Dialog Strategies

In each dialog turn a dialog system is confronted with a set of possible actions to execute. An optimal dialog strategy is the one which has the highest expected utility. Three major questions arise: what is the performance metric (how to quantify the success of a system), what are the necessary features and guidelines to consider in deciding the optimal strategy, and how does the system fuse all the features to come up with an optimal strategy?

In [24, 23] PARADISE was proposed as a general framework for evaluating a dialog system. Essentially the evaluation targets the relationship between *the user satisfaction* vs. *the task success* and *costs*. A multivariate linear regression is then conducted to find the performance function. The authors showed that the trained function has predictive power over the unseen dialogs, and therefore can be used for selecting the best dialog strategy. In [22] a reinforcement learning technique Q-learning is adopted to propagate back the estimated performance to previous dialog states in order to select the optimal strategy.

One interesting note to PARADISE is the proposition that the performance function is actually a linear combination of the various cost and success variables. It would be interesting to further investigate the alternatives, such as Bayesian networks, which are capable of capturing the dependencies between variables.

For the second major question of selecting optimal strategies, Gricean Maxims are often cited as one of the theoretical justifications of the approaches. Based on the Maxims in [4] an extensive list of principles of cooperative system dialog was derived from the analyses of a Wizard of Oz dialog corpus (collected by simulating a dialog system using human operators), and tested against a corpus collected from the interactions of the users and an actual dialog system. The refined principles can then be used as guidelines during the design and testing phases of dialog system development. Another application of the Maxims was reported in [3], where the author posited that user deviations from expectations, thus breaking the cooperative principle, indicates that the user is initiating a dialog repair turn. A preliminary evaluation was conducted with DIALOGOS system and showed a promising transaction success rate 84%. The similar idea of "deviation from expectation" was also exploited in [21], where the initiation of a verification dialog is a function of parse costs, deviation from the contexts, and domain-dependent exceptions. The author also gave important definitions concerning selective verification: *under-verification* denotes the situations where verifications are necessary but not executed, and *over-verification* represents the opposite situations.

The last major question concerning dialog strategies is how to fuse multiple information sources to make an optimal decision. An obvious approach, Bayesian networks, was reported in [12, 13, 14]. The architecture Quartet consists of two main modules: the maintenance module for handling uncertainties of channel and signals levels, and the intention module for handling the uncertainties of the user intentions. The conversation control thus maintains both the intention status (the simulated belief of the user) and the maintenance status (the self-evaluation of the performance). If the two show any discrepancy the control then adjusts its behavior. To *ground* mutual understanding by asking clarification questions, the system picks the question with the maximal sum of expected utilities of the decisions associated with each possible answer (Value-of-information analysis, VOI). However, the structure of the Bayesian networks in Quartet were manually constructed, and it is not clear how to scale the VOI analysis up in infinite domains (such as clarification questions over numbers).

4 Conclusions

Although in the literature the research to address the inaccuracies of a spoken dialog system often falls into one of the three main categories discussed in this paper, the boundaries are by no means clear-cut ones. In improving the recognition results, we already discussed works utilizing linguistic knowledge sources. In a similar vein the problem of insufficient linguistic knowledge can be remedied by executing appropriate dialog strategies (i.e., asking questions) to augment the system on the fly. The interactions between different categories can be even iterated: the recognition can be incremental, so the first version of the result can be immediately sent to the linguistic processor, and the partial analyses can be fed back to the recognizer to focus the effort on the most confusing part of the utterance. Thus a general architecture to accommodate different components and allow flexible interactions is necessary.

References

[1] E. Brill, R. Florian, J. C. Henderson, and L. Mangu. Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling? In *Proceedings of COLING-ACL '98*, Québec, Canada, August 1998.

[2] L. Chase. Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition. In *Proceedings of EuroSpeech 97*, Rhodes, Greece, 1997.

[3] M. Danieli. On the Use of Expectation for Detecting and Repairing Human-Machine Miscommunication. In *Proceedings of AAAI Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication*, Portland OR, 1996.

[4] L. Dybkjaer, N. O. Bernsen, and H. Dybkjaer. Reducing Miscommunication in Spoken Human-Machine Dialogue. In *AAI Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication*, Portland OR, 1996.

[5] M. Gavaldà. Interactive Grammar Repair. In *Proceedings of the Workshop on Automated Acquisition of Syntax and Parsing of the 10th European Summer School in Logic, Language and Information (ESSLLI-1998)*, Saarbricken, German, August 1998.

[6] M. Gavaldà. Epiphenomenal Grammar Acquisition with GSG. In *Proceedings of the Workshop on Conversational Systems of the 6th Conference on Applied Natural Language Processing and the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000)*, Seattle, U.S.A, May 2000.

[7] M. Gavaldà. Soup: A Parser for Real-world Spontaneous Speech. In *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy, February 2000.

[8] A. Lavie and M. Tomita. Efficient Generalized LR Paring of Word Lattice. In *The Third Symposium on the Foundations of Artificial Intelligence (BISFAI-93)*, Ramat Gan, Israel, June 1993.

[9] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, Budapest, Hungary, September 1999.

[10] R. Moore, D. Appelt, J. Dowding, J. M. Gawron, and D. Moran. Combining Linguistic and Statistical Knowledge Sources in Natural Language Processing for ATIS. In *Proceedings of ARPA Spoken Language Systems Technology Workshop*, 1995.

[11] X. Mou and V. Zue. The Use of Dynamic Reliability Scoring in Speech Recognition. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP-2000)*, Beijing, China, October 2000.

[12] T. Paek and E. Horvitz. Conversation as Action Under Uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, Stanford, CA, June 2000.

[13] T. Paek and E. Horvitz. DeepListener: Harnessing Expected Utility to Guide Clarification Dialog in Spoken Language Systems. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP-2000)*, Beijing, China, October 2000.

[14] T. Paek, E. Horvitz, and E. Ringger. Continuous Listening for Unconstrained Spoken Dialog. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP-2000)*, Beijing, China, October 2000.

[15] M. Rayner, D. M. Carter, V. Digalakis, and P. Price. Combining Knowdge Sources to Recoder N-best Speech Hypothesis Lists. In *Proceedings of the ARPA HLT Meeting*, 1994.

[16] E. K. Ringger and J. F. Allen. A Fertility Channel Model for Post-Correction of Continuous Speech Recognition. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP-96)*, 1996.

[17] C. P. Rosé and A. Lavie. A Domain Independent Approach for Efficiently Interpreting Extragrammatical Utterances. *Journal of Natural Language Engineering*, 1998.

[18] R. San-Segun, B. Pellom, and W. Ward. Confidence Measures for Dialogue Management in the CU Communicator System. In *Proceedings of ICASSP 2000*, Istanbul, Turkey, 2000.

[19] A. Sankar and A. Gorin. Adaptive language acquisition in a multisensory device. In R. Mammone, editor, *Artificial Neural Networks for Speech and Vision*, pages 325–356. Chapman & Hall, 1993.

[20] A. Sankar and A. Gorin. *Artificial neural networks for speech and vision*, chapter Adaptive language acquisition in a multisensory device, pages 325–356. Chapman & Hall, 1993.

[21] R. W. Smith and S. A. Gordon. An Evaluation of Strategies for Selective Utterance Verification for Spoken Natural Language Dialog. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, pages 41–48, 1997.

[22] M. A. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[23] M. A. Walker, C. A. Kamm, and D. J. Litman. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems.*, 2000.

[24] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Evaluating Spoken Dialogue Agents with PARADISE: Two Case Studies. *Computer Speech and Language*, 12(3), 1998.