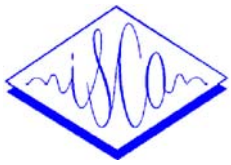

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Identifying Local Corrections in Human-Computer Dialogue

Gina-Anne Levow

University of Chicago
Chicago, IL USA
levow@cs.uchicago.edu

Abstract

Miscommunication in human-computer interaction is unavoidable, although speech recognition accuracy continues to improve. The perceived difficulty of correcting miscommunications has an even larger negative impact on assessments of system quality than does the absolute error rate. Therefore it is essential to improve error resolution capabilities in spoken language systems. While prior research has emphasized identifying the corrective status of an utterance, we focus in this paper on identifying the point of local correction. Users of spoken language systems often do not use specific syntactic structures or cue phrases to identify corrective intent or corrected content; most commonly a valid utterance is simply repeated, possibly slightly reworded. However, users do exploit prosodic cues to signal both presence and location of a correction. Using utterances from the 2000 and 2001 Communicator evaluation data collections, we build a boosted classifier to automatically identify the point of local correction in a corrective utterance. Exploiting the within sentence rank of prosodic cues including pitch maximum, pitch range, and intensity maximum, we distinguish locally corrected elements from other elements at 85.5% accuracy, a nearly 50% reduction in error rate over a naive majority class assignment.

1. Introduction

Although speech recognition accuracy continutes to improve, miscommunication in human-computer interaction is unavoidable. Furthermore, assessments of system quality [1, 2] have demonstrated that the difficulty of correcting a system error or misrecognition has a greater negative impact than does the absolute word or utterance misrecognition eate. Users may prefer more restrictive interfaces with frequent explicit confirmation to more conversational systems using only implicit confirmaiton because of the relative ease and immediacy with which errors can be detected, identified, and corrected. However, at the same time, users often bemoan the slow pace of system interaction. It is therefore highly desirable to improve and facilitate the human-computer error resolution process for spoken dialogue systems.

Error detection and resolution is performed smoothly and effectively in human-human interaction but presents some particular challenges in the case of human-computer dialogue. In particular, participants in a human-human dialogue may employ a wide range of lexical and syntactic cues to signal that a miscommunication has occured and what information was misinterpreted, ranging from cues phrases like "no I meant" to constructions like "it was X that I wanted." In contrast, because of the lexical and syntactic constraints on human-computer dialogue, these cues are either unavailable or rarely employed due to lack of confidence in suitable system interpretation. As a result, corrective utterances, when a user attempts to correct a system misrecognition - are often lexically and syntactically identical to other non-corrective inputs. Prior research [3, 4, 5] has identified and exploited a range of supra-segmental, acoustic-prosodic cues from pitch, intensity, duration, and pause to identify the corrective status of an utterance.

To enable more effective error resolution, it is necessary not only to identify that an error has occured but also to identify what error occured. In other words, we must determine not only that a correction is being made but also what, specifically, is being corrected. In many cases, such as rejection errors, the whole utterance in essence is being corrected. This circumstance is also likely to arise whenever the system gives inadequate feedback to its understanding or lack thereof. For example, if the user asks for "Duluth Minnesota" and the system misrecognizes it as "Dulles Minnesota" or some other invalid city state combination, reprompting with "Please say the city or state" does not give the user any information about the specific form of the misrecognition that should be addressed. The user's corrective utterance is thus likely to remain globally corrective, since the user has no information about what, if anything, the system thought it heard.

However, in some cases, the system provides enough information for the user to identify specifically what portion of their original input was misrecognized. The frequency of such cases depends on many factors including a system's dialogue strategy, underlying recognition accuracy, and rejection threshold. The user then has enough information to perform a more focused repair to identify the portion of the misrecognized utterance. We would like the system to be able to exploit any cues in the user's utterance to the location of the misrecognition in order to improve error resolution capabilities. By identifying the position of such a local corrective effort, we can identify where system should look for misinterpretation. Furthermore, other portions of the utterance can then be construed as successfully recognized. As few syntactic or lexical cues are available or exploited by users in the restrictive context of human-computer dialogue, we choose to exploit prosodic cues to identify the local focus of corrections in user utterances.

2. Background

Prior work has focused primarily on the identification of spoken corrections of system recognition errors. Work in English, German, and Swedish and other languages by [6, 7, 8, 9] has identified increases in utterance and pause duration as significant differences between original inputs and repeat corrections in live systems and in Wizard-of-Oz studies. Other prosodic features such as pitch and amplitude are less uniformly associated across languages, though they may in some cases provide significant contrasts, such as the decrease in pitch minimum

found in English corrections. [3, 4, 5] have demonstrated that these prosodic cues, possibly in conjunction with speech recognizer confidence scores, can be used to train machine learning classifiers to recognize corrections in human-computer interactions.

This task and these associated cues are also similar to those exploited for identification of self-repairs. Self-repairs tend to be accented or acoustically prominent, though this behavior is inconsistent.[10] While lexical overlap was found to be a useful cue, approaches [11, 12] exploiting prosodic cues alone or in conjunction with lexical information have also shown utility. These studies found that the repair region was characterized by increases in duration, pitch, and amplitude, though the magnitude of the latter changes was quite small.

[13] compared global and focal corrective adaptations in Wizard-of-Oz studies, where focal corrections involved a single misrecognized word either acoustically similar or not. For the position of focal correction, they found significant increases in duration, preceding pause, pitch maximum[1] and range, and intensity. For non-focal words, smaller increases in word and pause duration were observed, in corrective utterance relative to original inputs. These findings argue that focal corrections are distinguished both from original inputs and from non-focal regions within the repeat corrections themselves. However, little research has been done on automatically identifying the specific location of a local correction.

Next we will describe the data collection from which we drew our examples and precisely define the notion of a local correction. We will also characterize both our experimental data set, the analyses performed, and some acoustic contrasts found in local corrections, distinct from the remainder of the corrective utterance. Next we describe the feature set used to train the classifier, the basic classification framework, and the results of our experiments. We conclude with a discussion of these results and plans for future work.

3. Experimental Data

We briefly describe the corpus from which we extracted our examples and explain the criteria for selecting our local corrections.

3.1. Communicator Corpus

We used the 2000 and 2001 Communicator evaluation corpora for our collection and as a source of local corrections. This corpus includes recorded telephone interactions from more than 100 users using systems developed at nine different sites. The Communicator systems provided speech-only interfaces to travel planning functions to accomplish a set of tasks involving air travel, car rental, and hotel booking requests. Over the two years, more that 160 hours of audio interactions were recorded, digitized and transcribed. Unsurprisingly, the users encountered speech recognizer errors and engaged in error repair activities with the systems.

3.2. Local Corrections

From this large collection of interactions, we extracted a subset of utterances that exhibited clear local corrective behavior. Specifically we identified those utterances that immediately followed a speech recognizer misrecognition error, where the system's feedback to the user clearly indicated that a misrecognition had occurred and further clearly indicated a single word that required correction. We identified candidate utterances based purely on the transcripts and log files provided with the Communicator corpora. For lexically matched cases, we automatically chose instances where two sequential user input reference transcriptions matched word-for-word, but the speech recognizer hypothesis for the first differed. We then manually checked to see that the intervening system utterance clearly indicated a single word error. An example of such an interaction appears below. We refer to these cases as "lexically matched" or "matched" corrections.

User: May sixteen
System: I may have misheard you
saying the 6th of May.
Please be aware that the return date
must be later than the departure date.
User: May sixteen

While lexically matched corrections are quite common in the corpus, those involving both a single error and, more problematically, sufficient system feedback to identify the specific misrecognition are relatively rare. In order to obtain a slightly larger data set, we included both literal repeat corrections where the original input and the repeat correction were lexically matched and those which involved some rephrasing as below. In these cases, in the automatic candidate selection phase, we chose pairs in which the NIST sclite recognition scoring software indicated a substitution error and where a misrecognized word in the first utterance appeared in the second. The candidates were then manually filtered as described above. We refer to these cases as "lexically unmatched" or "unmatched" corrections.

User: October fifteenth
System: Okay, leaving on
October fifth. At what time?
User: The fifteenth of October

The lexically matched data facilitates direct word-by-word comparison of original and repeat correction utterances. We also relaxed the restriction the correction be the first in a possible error spiral sequence. This correction selection process yielded 36 lexically matched corrective utterances and 21 lexically unmatched corrective utterances exhibiting clear local corrections. These utterances included 57 words that functioned as local corrections from a total of 200 words in all corrective utterances, or 28.5% corrective words. Thus, a baseline most common class classification would have 71.5% accuracy.

4. Acoustic Features and Analysis

Following prior work, we consider prosodic features including duration, pitch, intensity, and pause. We employ the freely available Praat [14] acoustic analysis toolkit to compute pitch and intensity measures. We used the "To Pitch" and "To Intensity" functions to extract the corresponding feature values. We then apply a 5-point median filter to smooth out small variations in the pitch and intensity tracks. To provide comparability across speakers and call channels, all values are normalized on a per-utterance basis, computed as $\frac{val-mean}{mean}$, where, for example, val is the current observed pitch value and $mean$ is the per-utterance mean pitch. Forced alignment with the provided reference transcription, using the University of Colorado's Sonic speech recognizer, yields word boundary and thus duration and pause information. Per-word duration normalization is computed as $\frac{val-mean}{stddev}$, where the mean and standard deviation val-

[1] The increase in maximum applied to non-final words only. Final words exhibited a corresponding decrease in minimum pitch.

ues for the duration are based on phoneme duration values from ATIS (Air Travel Information System) data[15].

For duration, we find that the locally corrected word increases relative to its original counterpart. This increase is highly significant at $p < 0.005$. In contrast, there is no significant increase in duration overall for the other words in the utterance. No other changes in pitch or intensity for locally corrected words reached significance. However, these contrasts are known to wane during the course of error correction sequences, and our corrective pairs may be part of longer corrective sequences. In addition, such contrasts may have been employed differently across subjects.

5. Classification

We first describe the feature set provided for training and testing our classifier. We then describe the basic classifier framework and present our results.

5.1. Classifier Feature Set

For classification, we included features for normalized duration, pitch, intensity, and preceding pause for each word in the utterance. We computed the maximum and average values for normalized pitch and intensity. We also computed the pitch range, as normalized pitch maximum minus normalized pitch minimum. For each of these features for each word, we computed its corresponding within utterance rank - both absolute and normalized by sentence length. We consider the lexically matched data set, the unmatched set, and all data together. For each word, all these pairs of rank and value features plus position in utterance and its locally corrective status form the labeled training and test feature vectors for classification.

5.2. Classification Framework

We employed the freely available Boostexter [16] implementation of a boosted classification system. The boosted classification employs a weighted combination of weak learners to improve classification. Varying the number of rounds of training and the reweighting of the set of weak learners enabled us to avoid overfitting to training data, which had been problematic for decision trees given the data set size. The classifier in addition can provide information as to the features and thresholds employed by the learners at each round of training. We performed 5-way cross-validation, training on four-fifths of the data and testing on the remainder. We present the average of these results.

5.3. Classification Results

We obtain an average overall classification rate of 85.5%. This result improves over the baseline majority class assignment accuracy of 71.5% This rate represents a 50% reduction in error over the baseline. Considering only the lexically matched corrections, we achieve a classification accuracy of 81.25%, relative to a 59% baseline; for the lexically unmatched corrections, we achieve an accuracy of 87%, here relative to an 80% baseline. The difference in baselines is related to differences in utterance length. Unsurprisingly, exact lexically matched repetitions tend to be shorter, averaging between two and three words for those with focal corrections, while unmatched corrections vary more widely in length, here ranging from two to nine words in length. Overall they approach a one-half reduction in error.

Clearly these prosodic cues provide information to distinguish locally corrected words from the remainder of the utterance. However, it is important to note that our best classification accuracy was achieved when only rank-based information was provided to the classifier. Including the normalized values themselves in addition to the ranks increased the error rate to 20.5% overall. Furthermore, using the normalized values alone, without direct access to rank information led to performance at or even below the baseline. This contrast indicates that local corrective adaptations are clearly made relative to the utterance context in which they occur, rather than under some absolute degree of accentuation. Furthermore, the normalized values themselves introduce sufficient noise and variation to disrupt classification when employed in conjunction with the rank based information.

We inspected the features and thresholds selected for the weak learners during the Boostexter training process. We find that pitch range rank plays a particularly important role. In fact, a classifier using only pitch range rank achieves close to the best classification accuracy on the non-lexically matched subset. Likewise pitch maximum rank and intensity maximum rank are employed by the weak learners. Surprisingly, given its significant increase over non-corrective utterances, duration plays a less prominent role in classifiers. Finally word position also plays a role in classification, reflecting both a trend for local corrections to occur toward the end of the utterance - as a reflection of the global theme/rheme structure of the sentence as well as local syntactic constraints - and interactions between prosodic behavior and sentence position.

6. Discussion and Conclusion

Using utterances from the 2000 and 2001 Communicator evaluations, we have shown that a classifier, in this case, Boostexter, can be trained to automatically identify the position of a local correction in a spoken correction. We find that prosodic cues such as pitch and intensity play a particularly important role in achieving an 85.5% accuracy on this task, an improvement near 50% over a baseline majority class classification accuracy of 71.5%.

Interestingly, we find that it is not the value of these features that directly indicates corrective status, but their rank relative to other positions in the utterance. Furthermore, directly employing the values themselves in the classification process degrades performance relative rank-based features alone. This use of relatively large pitch range, pitch maximum or maximum intensity as some of the most prominent features in identifying the position of correction is consistent with the characteristics of accent and prominence. In general, prominent elements have wider pitch range, higher pitch, greater intensity and greater duration relative to both their context and the non-prominent form of the same word.

The lesser role of durational features, in spite of their statistically significant increases over their preceding utterance durations, may be attributed to the fact that many other words in the utterance increase in duration as well, due to the general effects of hyper-articulation in spoken corrections. [13] observed significant increases in duration for both focal and non-focal positions in corrections, although the magnitude of the change was greater for the focal elements. Changes in pitch range or intensity are less strongly associated with global corrective adaptations, and thus more effectively distinguish these local corrections, whereas changes in duration are heavily employed in both functions. These experiments demonstrate the relationship between prosodic cues and local corrective adaptations as part

of the general marking of prominence. They also provide additional insight into the interaction of global and focal corrective adaptations.

In future work, we hope to generalize this approach beyond single word corrections to phrasal and multi-position corrections. We also plan to consider more complex human-human corrective interactions where syntactic and lexical cues play a significant role in identifying corrections. In this process we hope to gain a better understanding of the integration of prosodic and other linguistic cues for detection and correction of miscommunications.

7. References

[1] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfcation," in *Proceedings of the DARPA Speech and Language Technology Workshop*. Morgan Kaufman Publishers: San Mateo, CA, 1992, pp. 49–54.

[2] M. Walker, R. Passonneau, and J. Boland, "Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems," in *Proceedings of ACL 2001*, 2001.

[3] G.-A. Levow, "Characterizing and recognizing spoken corrections in human-computer dialogue," in *Proceedings of COLING-ACL '98*, 1998.

[4] M. Swerts, J. Hirschberg, and D. Litman, "Corrections in spoken dialogue systems," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP'00)*, 2000, pp. 615–619. [Online]. Available: citeseer.nj.nec.com/swerts00corrections.html

[5] K. Kirchhoff, "A comparison of classification techniques for the automatic detection of error corrections in human-computer dialogues," in *Proceedings of the NAACL Workshop on Adaptation in Dialogue Systems*, 2001.

[6] S. Oviatt, G. Levow, M. MacEachern, and K. Kuhn, "Modeling hyperarticulate speech during human-computer error resolution," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 2, University of Delaware and A.I. duPont Instit., 1996, pp. 801–804.

[7] K. Fischer, "Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces," in *HCI International '99*, august 1999.

[8] H. Pirker, G. Loderer, and H. Trost, "Thus spoke the user to the wizard," in *EUROPSEECH-99*, 1999.

[9] L. Bell and J. Gustafson, "Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech," in *ICPhS '99*, 1999.

[10] W. Levelt and A. Cutler, "Prosodic marking in speech repair," *Journal of Semantics*, vol. 2, no. 2, pp. 205–217, 1983.

[11] E. Shriberg, R. Bates, and A. Stolcke, "A prosody-only decision-tree model for disfluency detection," in *Eurospeech '97*, 1997.

[12] C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *Journal of the Acoustic Society of America*, vol. 95, no. 3, pp. 1603–1616, 1994.

[13] S. Oviatt, G.-A. Levow, M. MacEachern, and E. Moreton, "Modeling global and focal hyperarticulation during human-computer error resolution," *Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 1–19, 1998.

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[15] G. Chung and S. Seneff, "Hierarchical modelling for speech recognition using the ANGIE framework," in *Proceedings of Eurospeech '97*, September 1997.

[16] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2–3, pp. 135–168, 2000.