EFFECTS OF WORD ERROR RATE IN THE DARPA COMMUNICATOR DATA DURING 2000 AND 2001

Gregory A. Sanders, Audrey N. Le, John S. Garofolo

National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899–8940
{gregory.sanders|audrey.le|john.garofolo}@nist.gov

ABSTRACT

During 2000 and 2001 two large data collections were performed, with paid users. We analyze the effects of speech recognition accuracy, as measured by Word Error Rate (WER), on other metrics. Analysis shows a linear correlation between WER and the Task Completion metrics, and (unexpectedly) this relationship remains more or less linear even for quite high values of WER. The picture for User Satisfaction metrics is more complex, and a linear model derived from the data by using the PARADISE framework [1] is given by Walker et al. [2]. We present evidence suggesting a somewhat linear relationship between WER and User Satisfaction for WER less than 35% or 40% in 2001, compared to stronger correlations in 2000. Finally, we note that the size of effect of increasing WER on Task Completion (slope of the least-squares regression line) appears to be about half as large in 2001 as in 2000, which we attribute to improved strategies for accomplishing tasks despite speech recognition errors. We consider this to be an important accomplishment of the research groups who built the Communicator implementations.

1. INTRODUCTION

The objective of the DARPA Communicator program is to support rapid,cost-effective development of automated systems with advanced ability to carry on spoken dialogues with a user who performs a task, such as travel planning, by using the system. Some had hypothesized that the success or failure of such systems would hinge almost entirely on the ability to achieve adequate speech recognition accuracy. This paper discusses the effects of the accuracy of Automatic Speech Recognition (ASR), as measured by Word Error Rate (WER), on various other metrics in the 2000 and 2001 data collections that were performed as part of the DARPA Communicator program. We believe there are two crucial metrics that indicate success: ability of the user to successfully perform the user's intended task, and user satisfaction. As a result of experience with existing spoken dialogue systems, we know *a priori* that such relationships exist. We focused on the following three questions. (1) What is the nature of the relationships between WER and those two crucial metrics? For example, are the relationships linear and is a linear regression thus appropriate? (2) What level of WER is compatible with successful task performance? Does task success "go off a cliff " beyond some level of WER? (3) What level of WER is compatible with high user satisfaction? These are important questions because many obvious applications for spoken dialogue systems involve noisy environments and distant microphones, which are factors that substantially degrade speech recognition accuracy.

2. THE DATA COLLECTIONS

Both the 2000 and the 2001 data collections used paid subjects. All subjects in 2000 were to be native speakers of American English and all were to use standard corded telephones. The subjects in 2001 were all English speakers but were not required to be native speakers of English, and the telephones in 2001 were allowed to include cellular phones and speaker phones. Thus, the subjects and the channel characteristics were more challenging in 2001.

There were nine Communicator systems participating in the 2000 data collection. The experimental design called for 72 subjects to each make nine calls, one to each of the systems. Subjects did seven hypothetical airline trip scenarios in their first seven calls, followed by two calls in which they were to plan real air trips for business and pleasure. Thus, the 2000 data collection was a within-subject design. Each subject made his/her calls over a period of just a few days.

Eight of the nine Communicator systems participated in the 2001 data collection, and the developers had the opportunity and funding to do significant research and development work to improve their systems between the 2000 and 2001 data collections. All data for 2000 reported in this paper is for only the eight systems that also participated in 2001.

The 2001 data collection involved about 200 subjects who made analyzable calls; all were frequent travelers. Some subjects were assigned to make four calls, all involving planning real trips. 77 of these subjects completed four analyzable calls. Other subjects were assigned to make ten calls, with the first four and last two being hypothetical travel scenarios and the fifth through eighth calls planning real trips. 69 of these subjects completed 10 analyzable calls. Two of the hypothetical scenarios were directly comparable to those used in 2000; the other four were more complex, involving three legs and/or rental cars and hotels. The 2001 experimental design assigned each subject to just one system, which the subject used repeatedly. Thus, the 2001 data collection was a within-system design, with the subject having the opportunity to learn (and exploit or avoid) the characteristics of a particular system. In 2001, each subject had access to the system for a period of a few months, and thus (if spreading his/her calls out

over a long time) had a chance between calls to forget previously acquired expertise about the system being used. Further details of the data collection can be found in the two papers by M. A. Walker et al., in the proceedings of this ICSLP 2002 conference

3. THE DATA WE ANALYZED

As one might expect, some calls had defects in the automatically generated transcripts/logfiles or in the user questionnaires that precluded calculation of the values of the relevant metrics. There were 133 such calls in 2001 (9.7% of the calls in 2001).

In addition, because the work reported in this paper focuses on the effects of speech recognition accuracy, it was desirable to identify calls for which the metrics would not have an opportunity to reflect the effects of WER. In particular, in both data collections, some sessions terminated with system crashes or suffered other unrelated failures very early in the conversation. We experimented with various schemes for picking out such calls by hand, but ultimately determined that a reasonable approximation could be obtained by excluding calls that were shorter than the shortest call that successfully completed an air travel planning task. In 2000, this meant excluding calls shorter than 6 user turns or 16 user words. There were 32 such calls in 2000 (5% of the calls), and there were 23 such calls in 2001 (1.7% of the calls). These suspiciously short calls were deemed defective and excluded. We have examined the transcripts of these calls and believe that deeming them defective is generally reasonable. Doing this increases the strength of some correlations but does not change the conclusions drawn.

4. RESULTS

Our presentation of these results is organized by metric. All subjects filled out a user questionnaire immediately after each call, giving their impressions of the call. In discussing the results of linear regressions, we say that an effect is significant if we reject the null hypothesis that the slope of the least-squares regression line is zero, and state significance levels for the effect on the basis of the significance level of the F statistic.

4.1. Speech recognition accuracy improved in 2001

Figure 1 shows the range of the middle 50% of the WER data, along with the median. As can be seen in Figure 1, WER decreased on average in 2001, from a median value of 24.2% to a median value of 17.3%, and became less variable. These improvements occurred despite the more challenging conditions of having some non-native speakers of English and some cellular or speaker phones in the 2001 calls. This improvement was also apparent in the WER results for the majority of the Communicator implementations that participated in both the 2000 and 2001 data collections.

In 2000, 95% of the analyzable calls had WER of 66.7% or less, with the data becoming somewhat sparse above WER of 55% (quite sparse above WER of 75%). In 2001, 95% of the analyzable calls had WER of 60.6% or less. The 2001 data becomes sparse at about the same level of WER as in 2000. Therefore, all un-smoothed data in the graphs in this paper corresponding to WER above 75% is based on relatively few calls.

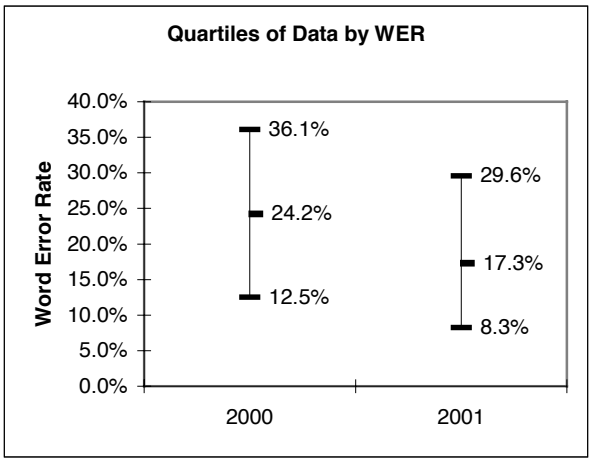

Figure 1: Word Error Rate Improved Between 2000 and 2001

4.2. The effect of WER on Task Completion metrics

The Communicator community has focused on User Satisfaction as a metric to analyze and optimize, but Task Completion is at least equally important. There are two task completion metrics: a subjective one from the user questionnaire and an objective one based on transcript annotation by a human expert. We call the subjective metric Perceived Task Completion (PTC). This item is a yes/no item on the user questionnaire that asked the user to state whether the travel planning task for the call had been successfully completed. The objective metric based on transcript analysis is called Annotated Task Completion (ATC). In 2001 we asked more or less the same PTC question as in 2000 but expanded the "success" choices to allow users to say that they had completed just airline reservations, or that they had completed air plus hotel and/or car reservations. Unfortunately, the questionnaire instructions were faulty for the first several weeks of the data collection. We have therefore created a CompositePTC metric by substituting the ATC values for the calls that occurred before we fixed the instructions. In general, correlations between ATC and other metrics such as WER are stronger than correlations with PTC or CompositePTC, and we believe this suggests that ATC is a more accurate or more objective Task Completion metric than PTC or CompositePTC.

In the 2000 data, a simple linear regression of WER against PTC gives r-square = 0.14, and the least-squares regression line has a slope of -0.0084 (suggesting that a 10% increase in WER would tend to decrease PTC by 8.4%). This effect is statistically significant at the 99.9% confidence level (95% confidence limits of -0.010 through -0.0067 for the slope).

The CompositePTC metric in 2001, converted to a binary value, is intended to be comparable to PTC in 2000. In the 2001 data, the corresponding simple linear regression of WER against CompositePTC gives r-square = 0.069 and the least-squares regression line has a slope of -0.0047 (suggesting a 10% increase in WER would decrease CompositePTC by 4.7%). This effect is also significant at the 99.9% confidence level (95% confidence limits of -0.0057 through -0.0037 for the slope). The 95% confidence limits for the 2001 data do not even overlap with the 95% confidence limits for the 2000 data,

and we therefore conclude that that PTC was significantly less dependent on the quality of ASR in 2001 than in 2000.

As mentioned, we also have an objective Task Completion metric, ATC. The linear regression of WER with ATC for the 2000 data gives r-square = 0.12 with the least-squares regression line having slope -0.0091, which effect is significant at the 99.9% confidence level, and the 95% confidence interval for the slope is -0.0111 to -0.0072. The corresponding linear regression of WER with ATC for the 2001 data gives r-square = 0.11 with the least-squares regression line having a slope of -0.0067, and this effect is also significant at the 99.9% confidence level, with the 95% confidence interval for the slope being -0.0078 to -0.0056. As the slope value for neither year's ATC data is within the 95% confidence interval for the slope on the other year's data, we conclude that (like PTC) ATC was significantly less dependent on the quality of ASR in 2001 than in 2000, thus demonstrating improvement. See Figure 3, showing this improvement. Note also that at a 95% confidence level, the slope for ATC is steeper than the slope for PTC in both years, as is illustrated by Figure 2. (The graph for the 2000 data corresponding to the 2001 data in Figure 2 shows this same pattern). As can be seen in Figure 2, this difference between ATC and CompositePTC lies more in the calls with higher WER, and we hypothesize that CompositePTC responses at higher WER levels are confounded with other (unknown) factors.

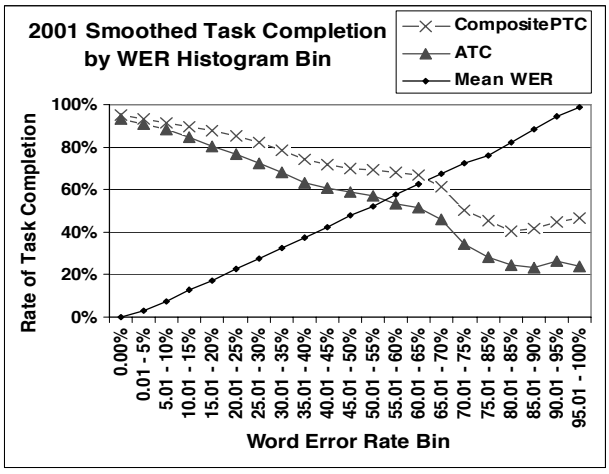

Figure 2: Task Completion decreases with increasing WER

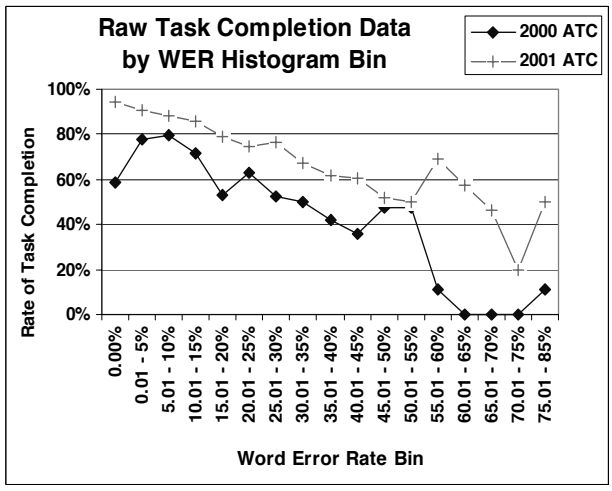

Figure 3: WER vs. Task Completion in 2000 and 20001

Contrary to our expectations, Task Completion appears to degrade linearly even to high levels of WER (see Figure 2). In interpreting Figure 3, the reader should however keep in mind that the underlying data is quite sparse above WER of 75%.

4.3. The effect of WER on User Satisfaction metrics

For purposes of characterizing the effect of WER on User Satisfaction, we have chosen to use only one item from the User Questionnaire. This is a Likert style item that asks for degree of agreement (Completely Agree through Completely Disagree) with the statement, "Based on my experience in this conversation using the system to get travel information, I would like to use this system regularly." We used only this one item because other items ask about factors that cannot be affected by WER. The results shown in Figure 4, a value of 5 on the y-axis corresponds to a response of Completely Agree (the most favorable response), and a value of 1 corresponds to Completely Disagree. A value of 3 is neutral. We note that at least in 2001, the responses hovered around neutral until WER exceeded 35%. A much more detailed model for User Satisfaction can be found in the two papers by M. A. Walker et al., in the proceedings.

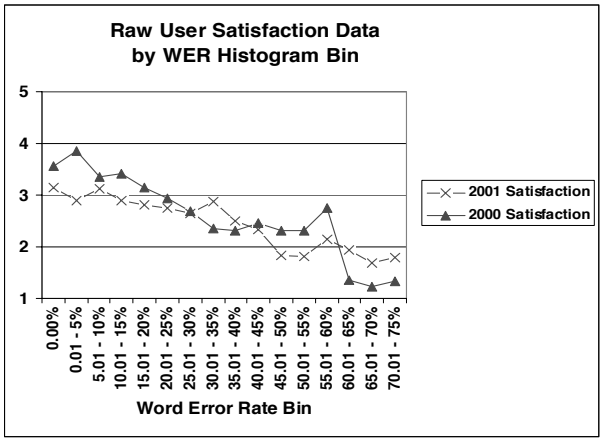

Figure 4: WER vs. User Satisfaction in 2000 and 2001

In 2001, WER and User Satisfaction were correlated only weakly and only for WER less than 50%. Most of the effect lies in the fact that the Completely Disagree responses to the User Satisfaction item on the User Questionnaire have a WER distribution that is a little higher on average than the other four (more favorable) responses. But 75% of even the calls that got Completely Disagree responses had WER of 40% or better, and for the 2001 data, a linear regression on the WER range 0.01% through 40% (n=786 calls), for WER predicting User Satisfaction, gives r-square=0.012, with slope (see Figure 4) of -0.018 (for the slope, F=7.34, with significance 0.007). The effect was greater in the 2000, but in 2000 task success was also more dependent on WER. We conclude overall, that any useful explanatory model for user satisfaction does not center on WER.

4.4. The effect of WER on Efficiency metrics

Usability studies traditionally construe efficiency as important. We have two principal efficiency metrics for the Communicator data: Time On Task (TOT) and User Words On Task (UWOT). We regard these as less important than Task Completion or User Satisfaction, but they illuminate the effects of WER on the conversations between systems and users.

The values of these two metrics are highly correlated, but Time On Task seems to be a more important determinant of user satisfaction. What we find particularly interesting about the relationship between these efficiency metrics, and WER is that at least in 2001 they were correlated with WER for WER rates up through about 35%, but then considerably more variable above WER of 35% (on multiple measures of variability). See Figure 5, in which we show the TOT and UWOT efficiency metrics averaged together, along with the other metrics we have been discussing (all rescaled to fit on one graph). In this graph, more efficient performance of the task is shown by lesser values on the y-axis: thus all three lines on the graph show deteriorating performance with increasing WER. Statistically speaking, there is no correlation or statistical evidence of an effect between WER and the efficiency metrics for high WER values (unless you cherry-pick the 40% through 60% range).

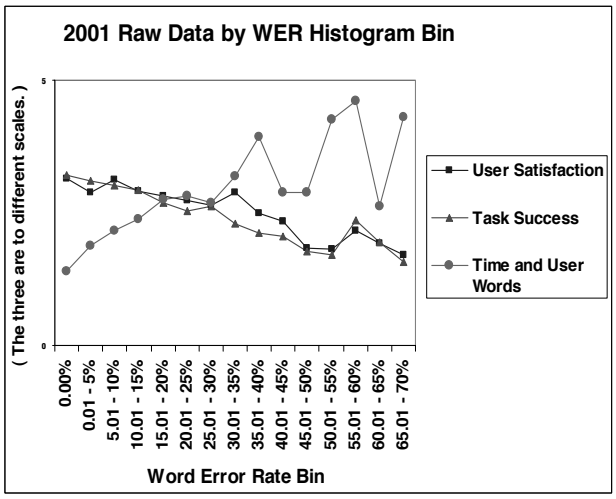

Figure 5: Efficiency metrics along with ATC and Satisfaction

4.5. Performance differs between systems even with WER held constant

Thus far, we have been building the case that WER is an important determinant of performance. Even so, WER is not the only factor, even for ATC. In Figure 6 we show a comparison among the eight systems that participated in the 2001 data collection, over the broad middle range of WER, with the data subset from each system carefully chosen/aligned to have equal mean WER and more or less equal variability. This is a grossly unfair comparison between systems in the sense that systems with lower WER ought to get the benefit of that aspect of their performance. But it does show clearly that even with WER made equal, performance on the metrics we have been discussing varies between systems. We believe much of the differences among systems revealed in Figure 6 can be attributed to the quality of dialogue handling in the respective systems. Elucidating the causes of the differences could be a topic for future work.

The Communicator program focused work on aspects of systems other than core speech recognition, per se. The fact that Figure 6 shows differences between systems that are not attributable to ASR suggests that was a worthwhile choice.

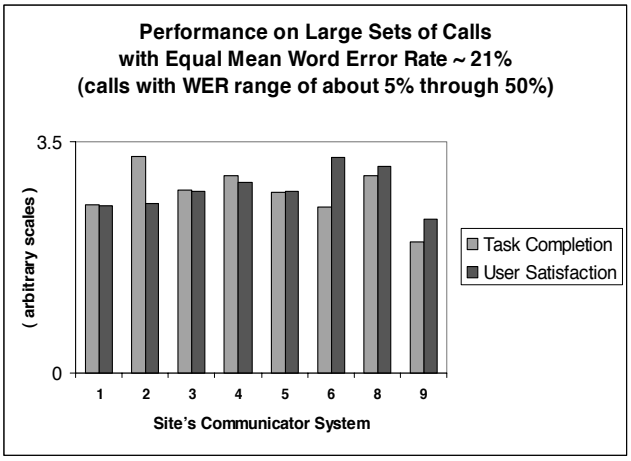

Figure 6: System performance differs even with WER equal

5. CONCLUSIONS

WER is an important factor in Task Success, which is more obvious when using the ATC metric than when using CompositePTC, and the effect seems to be linear even to high values of WER. The picture for User Satisfaction is more mixed and we think the best analysis is given in the two papers by M. A. Walker, et al. in the proceedings of this ISCLP 2002 conference. Efficiency metrics are probably most interesting as a determinant of User Satisfaction, and at high values of WER other factors probably swamp the effects of WER on efficiency. Finally, we note that values of WER above about 35% were associated with increased variability of multiple metrics in both 2000 and 2001, perhaps suggesting that a system's ability to deal with ASR errors becomes more influential above WER of 35%. In 2001, 80% of the calls had WER less than 35%, and 95% had WER less than 65%.

6. REFERENCES

[1] Walker, M.A., Kamm, C.A., and Litman, D.J., Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[2] Walker, M, Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S.,Narayanan, K., Papineni, B., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker, S. DARPA Communicator dialog travel planning systems: The June 2000 data collection. In *EUROSPEECH 2001*, 2001.