Measuring the Readability of Automatic Speech-to-Text Transcripts •

Douglas Jones [2], Florian Wolf [1], Edward Gibson [1], Elliott Williams [1], Evelina Fedorenko [1], Douglas Reynolds [2], Marc Zissman [2]

Department of Brain and Cognitive Sciences [1]
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
{egibson,fwolf,evelina9,elliottw}@mit.edu

Information Systems Technology Group [2]
MIT Lincoln Laboratory
Lexington, Massachusetts, USA
{daj,dar,maz}@ll.mit.edu

Abstract•

This paper reports initial results from a novel psycholinguistic study that measures the readability of several types of speech transcripts. We define a four-part figure of merit to measure readability: accuracy of answers to comprehension questions, reaction-time for passage reading, reaction-time for question answering and a subjective rating of passage difficulty. We present results from an experiment with 28 test subjects reading transcripts in four experimental conditions.

1. Introduction

A major goal of the new DARPA program for Effective, Affordable, Reusable Speech-to-Text (EARS) is to provide more readable automatic speech-to-text (STT) transcripts of news broadcasts and conversational telephone speech. [1] Ordinary STT transcripts are in single-case, lack punctuation, and include verbatim every word or word fragment that was spoken, including fillers such as "um", "uh", repeats, false starts, etc. The example in Figure 1 shows an error-free reference transcript that was created by trained human transcribers for the purpose of automatically scoring STT system output. The reference transcripts represent the theoretical upper bound for STT systems:

> yeah actually um i belong to a gym down here a gold's gym uh-huh and uh exercise i try to exercise five days a week um and i usually do that uh what type of exercising do you do in the gym

Figure 1: Reference Speech-to-Text Transcript **[STT_{ref}]**

The EARS program has defined speech metadata tags for annotating disfluent regions, structural units and speaker turns [2]. The intended benefits of the speech metadata fall into two categories (1) making the transcripts more readable for human readers and (2) improving automatic downstream processes that take the transcripts as input. Our focus here is on the human readers.

The transcript in Figure 2 shows a "transformed" reference transcript, which we refer to as an "XT" transcript. It is the maximally fluent rendering of the STT transcript and its associated metadata, given the imperfect nature of the input. The disfluent regions have been deleted and limited punctuation and capitalization have been added to render the text according to standard orthography.

> **A**: Yeah I belong to a gym down here. Gold's Gym. And I try to exercise five days a week. And I usually do that.
> **B**: What type of exercising do you do in the gym?

Figure 2: Reference Transformed Transcript **[XT_{ref}]**

We report here on preliminary work to test between the following hypotheses:

1. XT is more readable for human readers because disfluencies that made STT hard to read have been removed
2. XT is less readable for human readers because removing disfluencies deprives the reader of an important source of information

2. Types of transcript enhancement

Figure 3 shows our conceptual framework for organizing the role speech metadata plays in STT transformation. It begins with the sequence of temporally ordered words (as shown in Figure 1) at the top, which we've labeled STT–. More and more information is added that improves readability. The first step is to add speaker segmentation to break the blob of temporally ordered words into distinct turns labeled by speaker: we refer to this stage as STT+. The XT transcript at the last stage is the maximally fluent transcript.

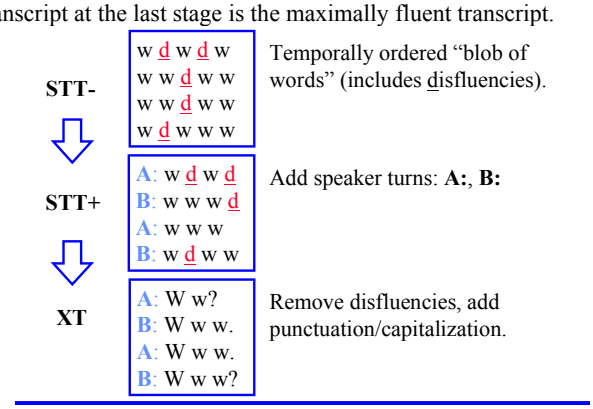

Figure 3. Three Stages of STT Enhancement

The distinction between STT– and STT+ is subtle but important in terms of technology development.[1] For four-

• This work is sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1] Therefore, for our experiments we have defined STT^{+} to use the reference speaker segmentation for all cases, even for STT

wire telephone speech, speaker segmentation is trivial if there is one talker per channel. But for broadcast news, assigning the correct labels to multiple speakers in the same channel is still a research challenge and a separate processing stage. We have treated this distinction with care in our study.

3. Transcript error

So far we have only discussed reference data which has been carefully prepared by expert human transcribers. Of course actual system output will contain error, both in the word recognition in the STT transcript and in the cleanup for the XT transcripts. Figure 4 shows system output for an STT+ transcript, which indicates speaker turns in the raw word recognition output:

A: actually uh i belong to a gym down here a gold jim uh i exercise so i tried exercise five days a week uh i usually do that
B: what took said can you imagine

Figure 4: System STT+ Transcript **[STT+$_{sys}$]**

This passage has a 50% word error rate (WER).[2] A reader who is expecting errors involving words that sound similar would not have too much trouble with errors like 'gold jim', and 'tried exercise'. But in the second turn, all but one of speaker B's words are incorrectly recognized, leaving that region extremely difficult to interpret. Furthermore, the errorful system output for the XT transcript in Figure 5 suggests that cleaning up the highly errorful transcript does not improve readability.[3]

A: I belong to a gym down here a gold jim. And I exercise so I tried exercise five days a week. And I usually do that.
B: What took said can you imagine?

Figure 5: System XT Transcript **[XT$_{sys}$]**

Therefore we also test the hypothesis of whether errorful system output is more difficult to read than reference transcripts.

system output, which may contain speaker labels of its own. For the raw STT output of broadcast news, we were not warranted in interpreting the STT output labels as true speaker segmentation labels because these labels may correspond to adapted speaker models rather than true speaker identification labels [3].

[2] Of the 42 words in the reference transcript, 23 were correctly recognized. There were 9 substitutions, 10 deletions and 2 insertions, giving 21 errors, or 50% error rate [4].

[3] We gratefully acknowledge the system output from one EARS research site which was prepared for a dry-run exercise. This output did not reflect the system's best capabilities, but served well as representative errorful data. The mean WER for the ten CT transcripts was 50%. Mean WER for the six BN transcripts was 20%. The reference experimental materials were drawn from the RT-03 dry run data from NIST [6].

4. Quantifying readability

We evaluate the readability of texts using some simple, standard measures from the psycholinguistic literature [5][4], quantifying the readability of a text in terms of four factors:

- participants' accuracy rates at answering the questions about the content of the text,
- the time it takes participants to answer questions,
- the time it takes participants to read the text and the questions,
- a subjective score that participants assign to the texts.

In particular, experimental participants are presented with short representative texts from the EARS Rich Transcript Evaluation dry-run data [6], consisting of four or five sentences – around 150-250 words from around 60 seconds of speech. On each trial, the participant is initially presented with the text as well as four questions about the informational content of the text. When s/he is finished reading the text and the questions, s/he presses a button, and the text and the questions disappear. The series of questions about the informational content of the text is then presented again to the participant, one at a time. Feedback (correct / incorrect) is provided to the participant after each question has been answered.[5] After these comprehension questions, the participant is asked to rate the text on a scale of 1-7 in terms of its difficulty (1-very easy, 7-very hard). See Figure 7 for an example.

5. Experimental design

5.1. Participants

32 participants from MIT and the surrounding community were paid for their participation. All were native speakers of English and were naïve as to the purpose of the study.

5.2. Materials

The materials for each experiment were formed from the initial EARS Rich Transcript Benchmark Tests transcript database, selecting 16 texts, 10 of which are conversational telephone (CT) speech, and 6 of which are broadcast news (BN) speech. The complete set of materials to be used in each experiment was formed by taking two short texts, between 150 and 250 words each, from each of the 16 texts, in order to obtain 32 total texts. The two short texts were taken from the larger texts in two locations: (1) Early: 150-250 words from the initial part of a text; (2) Later: 150-250 words from the middle/later part of a text. The sections of text were taken from regions where information was being transferred, avoiding regions of "chit-chat" in the CT texts. Figure 7 shows an example STT+$_{ref}$ text with questions.

[4] Although the psycholinguistics of reading is a mature field of scientific inquiry, we are aware of no previous work that addresses the readability of speech transcripts.

[5] The comprehension questions are designed primarily to ensure that the passages have been read. They present a challenge for the telephone conversations because of the chattiness of some of the passages. The reason we present the questions at the same time as the texts is to minimize memorization of text details.

Y: yeah actually um i belong to a gym down here gold's gym
X: uhhuh
Y: and uh exercise s- i try to exercise five days a week um and i usually do that uh now and then i'll i'll get it interrupted by work or something like that you know or just full of crazy hours or something you know i can't get out there but uh actually that's the kind of exercise i do i- i used to run a lot but uh when i got to be about twenty eight years old i just uh [noise] i- i don't know it was like some clock turned over or something [laugh] i just it was too sore and yeah didn't want to do it didn't find it enjoyable anymore and uh you know too much i guess to much pounding on the joints and stuff so i was sore a lot
X: what type of exercising do you do in the gym
Y: now i do strictly uh lifting and aerobics on like stair masters and uh stationary bikes
X: uhhuh good machines
Y: so yeah their i- i like the stair master it's uh really really easy on the knees but buil- seems to build up a lot of endurance so

Questions:

1) How many days per week does one of the speakers exercise?
- "1"
- "3"
- "5"
- "every second day"

2) What kind of exercise did one of the speakers do before he started going to the gym?
- "tennis"
- "weight-lifting"
- "bicycling"
- "running"

3) At what age did one of the speakers stop running?
- "34"
- "21"
- "28"
- "55"

4) What can sometimes keep one of the speakers from going to the gym?
- "traffic jams"
- "busy work schedule"
- "traveling"
- "seeing a movie"

Figure 7: Sample STT+$_{ref}$ text with questions

Each of the 32 texts had four versions: STT+$_{ref}$, XT$_{ref}$, STT+$_{sys}$ (error-full system output pre-clean-up) and XT$_{sys}$ (error-full system output post-clean-up). For the experiment, the conditions were balanced in a Latin Square design. Participants read either the early or the late part of each of the 16 original texts (10 CT + 6 BN) in one of the four conditions. Thus, participants read four texts per condition. The order of the texts in the experiment was pseudo-randomized separately for each participant, so that no two subsequent texts were from the same condition.

5.3. Procedure

Participants were seated in front of a computer screen. At the begin of each trial, the text appeared along with four comprehension questions, so that participants may use whatever reading strategy they like in order to answer the questions. They may choose to read the text first, and then answer the questions, or they may choose to read the questions first, and then scan back in the text for the answers. This experiment therefore more closely simulates a task in which people are scanning for particular kinds of information from a text. Participants pressed a button once they felt that they had read the text well enough in order to answer the questions about its contents. After participants pressed the button, they were asked to rate the difficulty of the text they had just read on a scale from 1 (very easy) to 7 (very hard). Then, the sequence of four questions about the content of the text appeared, one after another.

5.4. Predictions

If XT is easier to read than STT, XT should get a better (easier) difficulty rating than STT, the question answering accuracy should be higher for XT than for STT, and question answering time might be shorter for XT than for STT. The reading time for the text and questions might also be shorter. The reverse patterns of results would be expected if STT is easier than XT. If XT and STT are equally hard to read, no differences in our measures would be expected. There could also be an impact of system error on readability. If hand-coded reference texts are easier to understand than machine-coded system texts, reference texts should get a better difficulty rating, question answering performance should be better, and question answering time should be shorter for reference texts than for system texts.

6. Results

We excluded four of the 32 subjects who appeared not to be doing the task, each of whom had comprehension performance below 65%, compared with a mean of 87% and 73% for the worst performing of the remaining subjects. Reference texts were rated as subjectively less difficult than system texts (cf. Figure 6; ANOVA results being $F1(1,27)=147.0$, $p<.0001$; $F2(1,15)=23.6$, $p<.0005$)[6]. Furthermore, XT$_{ref}$ texts were less difficult than STT$_{ref}$ texts ($F1(1,27)=14.6$, $p<0.001$; $F2(1,15)=5.5$, $p<.05$). On the other hand, XT$_{sys}$ texts were rated as subjectively harder to understand than STT+$_{sys}$ texts ($F1(1,27)=3.6$, $p<0.07$; $F2(1,15)=4.5$, $p<0.06$). Questions on reference texts were answered better than questions on system texts ($F1(1,27)=12.1$, $p<.05$; $F2(1,15)=2.7$, $p<.12$), but there were no significant differences between XT and STT (cf. Figure 7; $Fs < 1$). The lack of difference between XT and STT+ was probably because comprehension accuracy rates were near ceiling. There were no significant differences for question answering times between XT and STT ($Fs < 1$). Because the XT and STT+ texts differed in length, a comparison in reaction time for the raw texts would not be very informative. In order to control for differences in text lengths, we also computed a residual text and question processing time (raw reaction time for reading the text and questions divided by the character length of the text and questions). Processing times for reference texts were shorter than for system texts (cf. Figure 9; $F1(1,27)=36.6$, $p<.0001$; $F2(1,15)=21.1$, $p<0.001$). But there was no significant difference between XT and STT texts ($Fs < 1$).

[6] Following standard practice in the psycholinguistic literature, we report analysis of variance using subjects as a random variable (F1) as well as items as a random variable (F2).

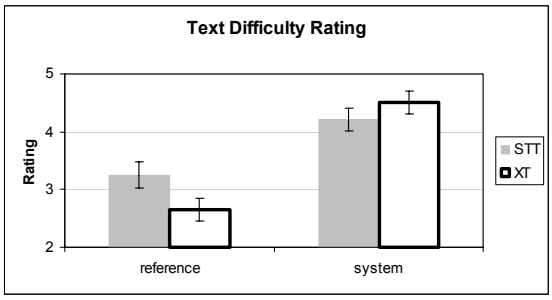

Figure 6: Perceived text difficulty ratings (1=easy; 7=hard)

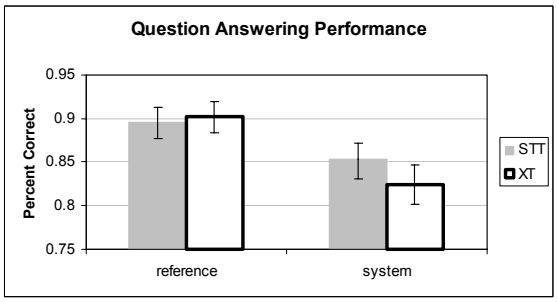

Figure 7: Question answering performance.

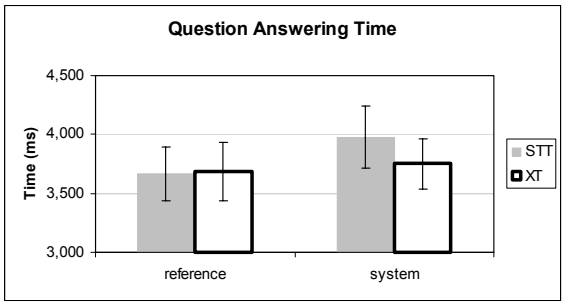

Figure 8: Question answering time.

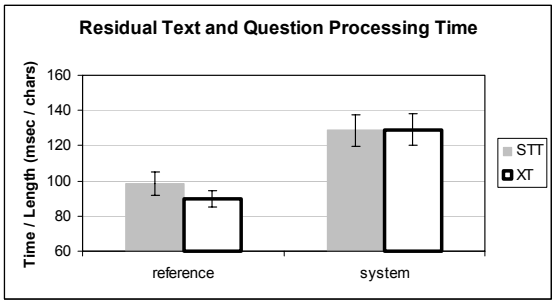

Figure 9: Residual text and question processing time.

7. Discussion

The text difficulty ratings suggest that reference texts are easier to read than system texts, regardless of whether they have been cleaned up or not. The difficulty ratings furthermore suggest that cleanup is helpful if it is done by hand for the reference texts; by contrast, system cleanup on system texts seems to decrease readability as measured by text difficulty ratings.

The results from question answering accuracy and time support the hypothesis that reference texts are easier to read or more suitable for information extraction than system texts, regardless of cleanup. Furthermore, the cleanup differences that appeared in the text difficulty ratings do not appear in the question answering data (accuracy or time). However, given the very high overall performance of participants in the experiment, XT-STT differences in the question answering data might be obscured by a ceiling effect. We are currently conducting experiments to address this issue.

8. Conclusion and future work

The framework we have defined captures the systematic subjective preference for XT texts over STT texts. It also captures objective differences between human transcribed reference texts versus errorful system texts in terms of question answering performance and text processing times. Our first experiment does not show objective differences for XT texts over STT texts, but there are reasons it is premature to conclude that there are not any. By co-presenting the questions and the text, we may have reached a ceiling effect by making the task too easy. We also do not know how much time is spent on the transcript as opposed to the questions. In future work we will present the questions separately from the text and we will test the hypothesis that the absence of speaker turns degrades readability.

Additionally, we will analyze the current results with more detailed information about the word error and cleanup. For example, we will perform correlational analyses on the texts between (1) the quantity of each clean-up operation / error type and (2) the readability scores for a text. For instance, if including periods between sentences improves readability, then adding periods should improve one or more readability scores. The results of the correlational analyses will help determine which kinds of word errors or cleanup are most important with respect to readability. We will also probe systematic differences between Broadcast News versus Conversational Telephone Speech.

9. References

[1] Wayne, C., *Effective, Affordable, Reusable Speech-to-Text (EARS)*. Official web site for DARPA/EARS Program. http://www.darpa.mil/iao/EARS.htm. 2003.

[2] Strassel, S., *Guidelines for RT-03 Transcription -- Version 2.2*. Linguistic Data Consortium, University of Pennsylvania. Manuscript. February 23, 2003

[3] Martin, A, and M. Przybocki. "The NIST 1999 Speaker Recognition Evaluation – An Overview". *Digital Signal Processing, **10***, 1-18. 2000.

[4] Fiscus, J., *Speech Recognition Scoring Toolkit (SCTK) Version 1.2c* http:// www.nist.gov / speech / tools. 2000.

[5] Kintsch, W., *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press. 1998.

[6] Garofolo, J., *Rich Transcript Spring 2003 Benchmark Tests*. http:// www.nist.gov /speech / tests / rt / rt2003