MEASURING HUMAN READABILITY OF MACHINE GENERATED TEXT: THREE CASE STUDIES IN SPEECH RECOGNITION AND MACHINE TRANSLATION •

Douglas Jones[1], Edward Gibson[2], Wade Shen[1],
Neil Granoien[3], Martha Herzog[3], Douglas Reynolds[1], Clifford Weinstein[1]

[1]MIT Lincoln Laboratory
[2]Department of Brain and Cognitive Sciences, MIT
[3]Defense Language Institute Foreign Language Center

Abstract

We present highlights from three experiments that test the readability of current state-of-the art system output from (1) an automated English speech-to-text system (2) a text-based Arabic-to-English machine translation system and (3) an audio-based Arabic-to-English MT process. We measure readability in terms of reaction time and passage comprehension in each case, applying standard psycholinguistic testing procedures and a modified version of the standard Defense Language Proficiency Test for Arabic called the DLPT*. We learned that: (1) subjects are slowed down about 25% when reading system STT output, (2) text-based MT systems enable an English speaker to pass Arabic Level 2 on the DLPT* and (3) audio-based MT systems do not enable English speakers to pass Arabic Level 2. We intend for these generic measures of readability to predict performance of more application-specific tasks.

1. INTRODUCTION

Research in Human Language Technology has made great progress in the past few years. Automatic speech-to-text (STT) systems are capable of producing English text transcripts of conversational telephone speech at 15.2% word error rate, a reduction of 53% over the past 5 years. Arabic to English machine translation (MT) systems are capable of producing English text output at BLEU scores [1] of 0.47, an increase of over 300% over the past 3 years. These measures of performance result from technology-centered evaluations. In this paper we outline a framework for addressing the question of how these remarkable gains are reflected in measures of effectiveness which show the impact of the technology on the effectiveness of human users in accomplishing specific tasks.

Our paper is organized as follows: we begin with an outline the overall goals, background and general framework for our experiments. We then show highlights from our three case studies which measure the effects of system errors on human readers. In the conclusion we mention future work and overall implications.

2. GOALS OF STUIDIES

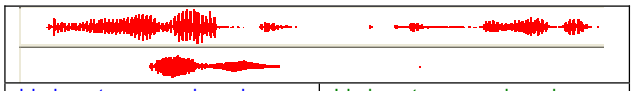
A major goal of current DoD-sponsored research in speech recognition is to provide more readable automatic STT transcripts of news broadcasts and conversational telephone speech. Figure 1 contrasts transcripts produced by an experimental STT system (left), which contains word recognition errors, with a human-produced reference transcript (right). The transcript on the right is the gold standard by which the system output is scored and serves as a target for technology research and development. The benefits of improving automatic speech transcripts fall into two categories (1) making the transcripts more readable for human readers and (2) improving automatic downstream processes that use these transcripts as input. In this discussion we focus on the human readers.

Likewise, a major goal of DoD-sponsored research in MT is to build automatic systems that translate Chinese and Arabic texts into English texts that can be used by English native readers to perform pertinent tasks. Figure 2 shows a fragment of an Arabic Level 3 text, rated using the Interagency Language Roundtable skill levels, and sample questions that a competent Arabic reader could be expected to answer. The translation on the left is produced by a state-of-the-art MT system; the translation on the right was produced by professional human translators.

In Figures 1 and 2, the intuition is very clear that the human transcripts and translations are easier to read. What is not clear is the level at which these technologies can enable their intended consumers (e.g. analysts) to perform real tasks. In order to quantify the effectiveness of these technologies and provide feedback for research programs such as DARPA EARS (Effective, Affordable Reusable Speech-to-text) and DARPA TIDES (Translingual Information Detection Extraction, and Summarization), we launched a project that applies rigorous psycholinguistic experimentation and government-standard proficiency evaluation techniques.

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

Our intent is for these generic measures of effectiveness for human readability to predict effectiveness on performing similar types of real-world tasks.



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 what took said can you imagine	I belong to a gym down here, a Gold's gym. I try to exercise five days a week, and I usually do that. Now and then, I'll get interrupted by work.

Figure 1: Speech-to-Text Transcript: System vs. Human

[1]رأى الأهرام -- [2]تبدو تصريحات وزير خارجية الحكومة العراقية المعينة, التي هدد فيها بإمكان السماح للقوات الأمريكية التي تحتل العراق بشن هجمات علي البلدان المجاورة للعراق, ردا علي دعمها لعمليات المقاومة العراقية. ...

Al-Ahram Al-view	Al-Ahram Opinion column
Foreign minister's statements appear to be designated the Iraqi government, which threatened the American forces, which could be allowed to occupy Iraq to launch attacks on neighbouring countries to Iraq in retaliation for its support for Iraqi resistance. ...	Declarations by the Minister of Foreign Affairs in the "appointed" Iraqi government threatening to allow the American forces that occupy Iraq to launch attacks on neighboring countries seems to be in response to their support of the Iraqi resistance. ...

Questions:
1. List one of the author's expectations for the new Iraqi regime.
2. Why does the author consider the minister's remarks dangerous for the Iraqi government?

Figure 2: Arabic to English: System vs. Human

3. GENERAL FRAMEWORK FOR EXPERIMENTS

We employed a general framework for experiments across different modalities and technology applications. We start with source materials in a variety of different formats: English audio signals from broadcast news, conversational telephone speech, foreign language texts from newspapers, etc. These materials are converted to standard English text (either by translation or transcription) and then presented to human subjects who are then asked to read and answer comprehension questions. We present texts that were generated by machines and those generated by humans. We then measure the human's ability to answer questions about the text and the speed at which they are able to process the text. With gold standard human processed texts, we expect a higher level of performance and lower reading times from human subjects. With experimental machine system output, errors may occur during transcription or translation, yielding texts that are garbled. We expect a lower level of performance with higher reading times.

Using results from both machine and human generated output it is then possible to quantify the degradation that these automatic translation and transcription techniques introduce, in a variety of test conditions (ranges of errors, types of transformations, etc.). In the next section we present highlights from three recent experiments.

4. THREE CASE STUDIES

In measuring the readability of texts produced by automated STT systems and MT systems, we use simple, standard measures from psycholinguistic research [6] and computer-based language testing [5] , quantifying the readability of a text in terms of three factors:

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

A common theme in the three case studies is that we have attempted to measure the technology in an application-neutral fashion. We have created a test bed of general purpose comprehension questions that stay fixed across a variety of experimental conditions. Application-specific experiments such as the work by Hirschberg and her colleagues [7] are very important in measuring the effects on end-users. In general, we draw from a range of delivery and timing options in our test protocols, from self-paced reading at the word level and at the line level, to more coarse-grained times for whole-passage reading and question answering. In the three case studies presented in this paper, we use whole-passage reading times.

4.1 English Speech-to-Text Transcripts

Our first case study investigates the readability of automated STT transcripts. As reported in [1] we conducted experiments using short representative texts from the EARS Rich Transcription Evaluation dry-run data, consisting of four or five sentences – around 150-250 words from around 60 seconds of speech. Using this data we compared subjects' performance with conventional, verbatim gold-standard reference transcripts and their errorful system-produced counterparts. All of these transcripts contained disfluencies. Moreover, they were not capitalized or punctuated according to conventional English orthography. The system word error rate ranged from 8.6% to 58.3%, with a mean WER of 30.4%. The participants were presented with the text as well as four questions about the informational content and their responses were timed. In order to control for differences in text lengths, we computed reaction time in terms of the subjects' residual text processing time [6] . We found that reaction time was slowed significantly, by about 25%, from 0.52 seconds per word to 0.65 seconds per word, as we see in Figure 3. Comprehension performance suffered from the relatively high word error

rate, but only modestly, as we see in Figure 3, with a mean accuracy drop to 85% from 90%. These and other details were reported in [1]

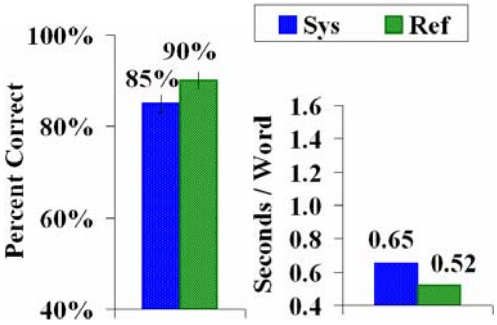


Figure 3: Conventional STT Transcript Readability

In related work, we investigated the effects on readability of various types of clean-up for conventional STT transcripts: adding punctuation, removing disfluencies, etc, both in the reference condition of human-produced transcripts as well as in conditions of errorful system output. The effects of transcript clean-up are more subtle than those of word error rate differences, but in general we observe modest improvements of readability for transcript clean-up in the 5-10% range for speed and accuracy. Details were presented at the Rich Transcription 2004 workshop [3] .

4.2 Arabic to English Text MT and the DLPT* Test

Our second case study highlights work from a recent experiment designed to test the utility of MT at varying levels of quality. In this experiment, educated English-native speakers answered questions from a machine translated version of a standardized Arabic language test (the Defense Language Proficiency Test or DLPT) specially constructed for testing MT quality. The DLPT is a high-stakes test that has been administered in the U.S. Defense Department for decades to determine the suitability of candidate linguists to perform real-world tasks that require foreign language proficiency. Our modification of the test, the DLPT*, substitutes machine-generated English output in place of the original foreign language materials. Our intent is that the DLPT* test will predict performance of English speakers performing real-world tasks with the machine-generated output in place of the original materials.

In our experiment, we compared MT results with professional reference for the purpose of determining the level of Arabic reading comprehension that current MT technology enables an English reader to achieve. Following DLPT practice, 70% comprehension at a given level and all levels below is required to pass a given level. The results of this experiment suggest that MT may enable subjects to achieve an Interagency Language Roundtable Level 2 performance, but is not yet adequate to achieve ILR Level 3. Reaction times are generally slower when reading MT output in comparison with human translations for each level. Across levels, subjects report different strategies: for the longer, more difficult Level 3 items, they typically look at the questions first and then scan for answers. For the shorter Level 2 items, they often read the passage first. The different ranges of reaction times reflect this type of variation in performing the task. These and other details have circulated informally in a project report for DARPA [4]

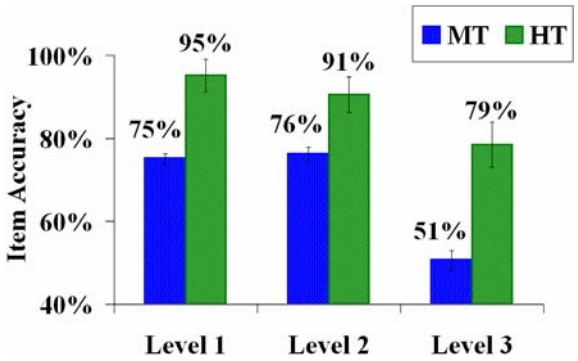


Figure 4: Arabic Text MT Comprehension

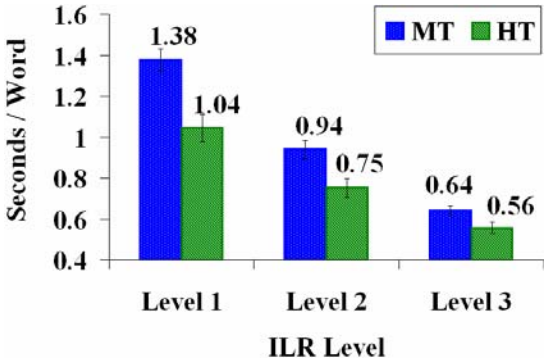


Figure 5: Arabic Text MT Reaction Time

4.3 Arabic to English Audio Machine Translation

Our third case study highlights preliminary results from an experiment currently in progress to test the utility of an Arabic STT system combined with an Arabic-to-English MT system, using the DLPT* test. Results from 22 Level 2 texts are shown in Figure 6. These materials are from the Arabic text MT experiment mentioned in the previous section. The passages were read out loud by a diverse set of speakers: two male speakers and three female speakers from Jordan, Morocco, Egypt and Sudan, producing around 33 minutes of Arabic audio. These materials were then processed by a state of the art Arabic STT system, operating at 27.3% word error rate. In addition to producing Arabic words, the system produced segmentation of sentence-like units. This output was in

turn sent to a state of the art Arabic-to-English MT system, operating at BLEU score of 0.28 (compared with BLEU scores of 0.37 for MT with reference text as input). The substantial reduction in performance can be attributed to the STT errors. These materials were then read by 32 human subjects taking the DLPT* test. In this experiment, we observed that Audio MT fails to pass Level 2, since it is below the passing threshold of 70%. The reading times for the Audio MT materials are slightly slower than for the human reference materials, as shown in Figure 6.

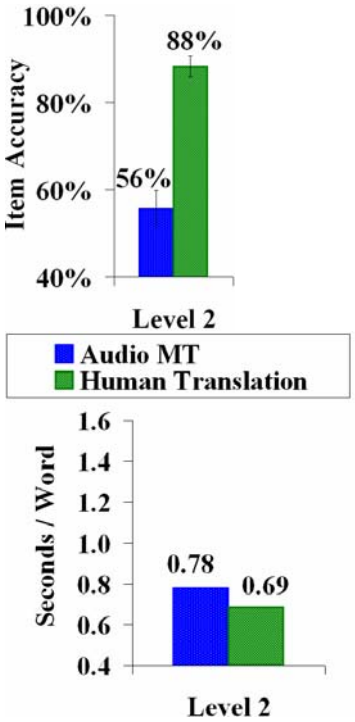


Figure 6: Arabic Level 2 Audio MT Readability

5. CONCLUSIONS AND FUTURE WORK

We believe that human subject experiments of the sort we discuss in this paper provide a useful tool for quantifying the effectiveness of technologies in downstream human processes. These experiments help us to quantify the intuition that errorful system output is harder to read than human-produced gold-standard data. Errorful speech transcripts slow people down and interfere with their abilities to answer comprehension questions – the modest effects on comprehension are likely due to the fact that the telephone conversations in the experiments not particularly rich in information content. For MT, we are able to provide general levels of expected utility using DLPT* tests – current state-of-the-art Text MT systems can enable English speakers to operate at Level 2, but not yet Level 3, and state-of-the art Audio MT systems have not yet achieved Level 2 capabilities.

In future work, we will investigate specific ways in which system errors influence readability at smaller levels of granularity. We will explore avenues for improving Audio MT by improved STT segmentation and related interface factors that may have an effect on human readability. An important area for future work is to study the relationship between the system measures of performance, such as WER and BLEU, with the measures of effectiveness, such as scores on the DLPT*. Another important area to explore is the relationship between scores on the DLPT* and the performance of various real-world tasks that the original DLPT test was intended to predict. Additionally, we will explore the application of this experimental procedure to other types of HLT technologies as well as other types of human language tasks.

6. ACKNOWLEDGMENTS

We wish to acknowledge several people who performed critical parts of the experiments: Hussny Ibrahim and Osaila ElKhatib at DLI, for scoring the Arabic DLPT* tests, as well as oversight by Deniz Bilgin. Charlene Chuang and Rina Patel administered the STT readability tests and the Arabic Text MT tests at MIT. John Tardelli was responsible for administering human subjects experiments for the DLPT* at ARCON. Chris Cieri, Tim Buckwalter, Mohamed Maamouri and Kevin Walker arranged for the Arabic texts to be read aloud at LDC and then prepared the gold-standard transcripts. Bing Xiang, Long Nguyen and John Makhoul at BBN processed the Arabic audio files for STT and automatic segmentation. Kevin Knight and Ignacio Thayer at USC/ISI processed the Arabic-to-English MT request.

REFERENCES

[1] Papineni, K., et al. 2001. "BLEU: a Method for Automatic Evaluation of Machine Translation" *IBM Computer Science Research Report RC22176 (W0109-022)* 9/ 17/2001.

[2] Jones, Douglas A., et al. 2003. "Measuring the readability of automatic speech-to-text transcripts", EUROSPEECH-*2003*, 1585-1588.

[3] Gibson, Edward, et al. 2004. Two New Experimental Protocols for Measuring STT Readability. 10/31/2004. Draft Report for DARPA/EARS/Rich Transcription 2004 Workshop.

[4] Granoien, Neil, et al. 2004. “Enabling English Speakers to Pass Level 3 on a Defense Language Proficiency Test for Arabic”. 10/29/04. Report for DARPA Pre-GALE Utility Study.

[5] Herzog, Martha. 2003. New Testing Approaches for New Languages. East Coast Organization of Language Testers.

[6] Just, M. A., P. A. Carpenter, et al. 1982. "Paradigms and processing in reading comprehension." *Journal of Experimental Psychology*: General 111: 228-238.

[7] Whittaker, S., et al. 2002. SCANMail: a voicemail interface that makes speech browsable, readable and searchable, *Computer Human Interaction 2002*, pp. 275-282.