Caption Accuracy Metrics: Solutions for Automatically Measuring Caption Quality

WGBH National Center for Accessible Media

CSUN Conference
March 16, 2011

In today's discussion…

- Background and status of caption quality issues
- Related current legislation
- Caption Accuracy Metrics project
  - scope and activities
  - participants
  - research, development and data analysis
  - outcomes

About the Caption Accuracy Metrics project

Funded by the U.S. Department of Education, National Institute on Disability & Rehabilitation Research

- Three year grant: October 2008 - September 2011
- Deliverables include:
  - Publication of an experimental ontology of error types
  - Publication of research into error capture capabilities of text mining software
  - Development of a software prototype application

About the Caption Accuracy Metrics project

This grant was awarded to the National Center for Accessible Media (NCAM) to:

Develop a prototype measuring tool that can analyze the quality of real-time captioning, developed with input from industry leaders, deaf education experts and the National Institute of Technology and Standards (NIST).

First.. how we got here, and progress gained…

Caption quality issues/current legislation: Larry Goldberg, Director of WGBH National Center for Accessible Media:

- Background & status of caption quality issues
- The 21st Century Communications and Video Accessibility Act

What the CC Metrics project is…and isn't

The project is:

- ... testing the ability of data-mining software to identify discrepancies between traditional stenocaption text and speech recognition text
- … generating a caption accuracy analysis of the data sets under review

The project is *not*:

- … attempting to use speech recognition technology for real-time captioning
- … publicly associating test data with specific TV networks, TV programs, or caption agencies

CC Metrics project outcome and benefits

Project outcome:

The project will develop a customized prototype that enables standardized, independent analysis of caption accuracy metrics.

Benefits:

- Provide standardized tool for caption accuracy measurement
- Improve ability of TV industry to monitor and maintain quality of captioning services
- Reduce need for caption viewers to document and advocate for better quality

CC Metrics project phases

- Identify standard error types
- Rank error types based on consumer feedback
- Create reports using manually-generated “ground truth” transcripts and text mining tools
- Substitute ASR transcripts and compare results

Technical review panel and other participants

Our goal: unite a variety of stakeholders in caption quality issues to inform project outcomes.

- *National Court Reporters Association*
- *Caption agencies*
- *Advisors from Gallaudet University and NTID*
- *Broadcast and cable networks*
- *Nuance*
- *NIST*
- MIT Computer Science & Artificial Intelligence Lab

CC Metrics Year 1 activities

- Created baseline data set of 18 real-time network news programs
- Measured accuracy using standard alignment tools from NIST
- Experimented with speech recognition transcripts for comparison
- Evaluated data/text mining software tools and options for customization

Caption errors

Three basic error types:

- Substitutions
- Deletions (Drops or Omissions)
- Insertions (Additions)

Sample error #1

Spoken:

>> THIS PROCESS WILL BE QUICK.

Caption:

>> THIS PROSWILLING QUICK.

Sample error #1

ALIGNED

>> THIS PROCESS WILL BE QUICK.

>> THIS PROSWILLING ** ** QUICK.**
S D D

3 errors: 1 substitution, 2 deletions

Sample error #2

Spoken:

>> SMOKING DEATH RATES HAVE CONTINUED TO INCREASE

Caption:

>> THE SMOKING DEATH RATE HAS INCREASED

Sample error #2

Aligned:

```
*** SMOKING DEATH RATES HAVE CONTINUED TO INCREASE
THE SMOKING DEATH RATE HAS  **********  **  INCREASED
I                       S     S     D         D   S
```

6 errors: 1 insertion, 3 substitutions, 2 deletions

6 errors out of 7 words = 85% error rate

Caption Metrics

Word Error Rate

$$\text{WER} = \frac{S + D + I}{N}$$

Basically, total errors divided by total number of words

CC Metrics Year 2 activities

- Refined error types and refine draft error ontology
- Designed, launched national consumer survey, began initial analysis
- Built recording capture station to amass larger data sets
- Further explored data mining and speech recognition software utilities and development options for customization
- Began work with Nuance to customize language processing tools and data analysis software
- Held technical panel meeting with industry representatives

Ontology of error types

- Break down three categories (S,D,I) in more detail
- Draws on NCRA guidelines: What is an error?
- Identified 17 types of errors
- Not addressing every possibility

Substitution errors (mild)

1. Singular/plural
2. Wrong tense
3. Substitute nominal (pronoun) for Proper Name
4. Punctuation
5. Split of compound word, contraction (OK)

Substitution errors (severe)

12. Nearly same sound (homophone)
13. Wrong word
14. Similar sound but steno
15. Garbled syllables, not words
16. Word boundary error
17. Transmission problems (e.g., garbling, white boxes, dropped letter pairs, etc.)

Insertion errors

6. Split of compound word (one word or both wrong)
7. Duplicate word or minor insertion
8. Word order (transposition)
9. Correction by steno

Most insertion errors are in combination with other error types.

Deletion errors

10. Dropped words: 1-2 (minor, aside)

11. Dropped words: 3+ (significant)

Context is key: a single word drop can be critical – or inconsequential.

Overview of work with Nuance

- Research whether text-based data mining and speech-to-text technologies can produce meaningful data about stenocaption accuracy
- Determine methods of using language processing tools to enable independent analysis of caption accuracy metrics
- Weight error severity with data from consumer survey
- Develop and refine software capabilities for automated caption accuracy system

Status of work with Nuance

- Complete test data set delivered to Nuance
- Sample “tagged” newscasts (errors manually defined)
- Demonstrated preliminary alignment and basic error identification (S, D, I) with subset of data
- Refined alignment and error identification with full data set
- Designed prototype architecture and error reports design

Consumer survey overview

- Developed survey to better understand from TV news caption viewers about types of errors that make dialogue hard to follow

- Broad national distribution of survey invitation:
  - Disability-focused listservs and blogs
  - NCAM's own extensive list of individual and organizational contacts
  - Many national and regional consumer advocacy groups redistributed invitation to their constituents

Consumer survey overview

- Specified survey was *only for people who use captions when viewing television news*
- Emphasized that goal of survey was to ask about the impact of different error types, *not to test caption-reading skills.*
- Launched for three weeks in spring 2010
- Respondents were able to fill out the survey in increments and return to the survey

Consumer survey demographics

- 351 respondents completed the survey
- 48 states represented
- Respondents self-identified as:
  - Deaf (50%)
  - Late deafened (12%)
  - Hard of hearing (29%)
  - Hearing (9%)
- 62% between 30-60 years old

74% indicated they watch one or more newscasts every day

Consumer survey design

- Survey consisted of three sections:
  - A: collected demographic info and TV viewing habits (respondents were asked to self-identify as Deaf, Late-Deafened, Hard of Hearing, or Hearing)
  - B: presented 15 actual caption examples as still-frames
  - C: presented 26 additional caption examples as text
- Examples correlated to every category of error type and severity from the error ontology
- Attempted to only show one error per example

Consumer survey design

Section B used still frames instead of video:

- to allow respondents to focus on the error
- to save response time and allow for inclusion of more examples

Consumer survey design

Response choices - for each caption example in section B, respondents were asked to choose one of the following:

- I do not notice an error
- The caption has an error but it does not bother me (minor error)
- The caption has an error that bothers me (major error)

Consumer survey design

- In section B, if respondents did not notice an error, they were presented with the next example.
- If they did notice an error, a follow-up question asked if/how the error would affect the respondent's understanding of the caption:

  - No, I understand the caption
  - Yes, it would somewhat affect my understanding
  - Yes, it would greatly affect my understanding
  - Yes, it would completely destroy my understanding

Consumer survey design

Section C used text to show the caption example alongside of what was spoken, with the error(s) underlined in the spoken word text:

Example of captions as they appeared onscreen:

THE RULES ARE VERY HARD
FOR ANY THIRD PARTY TO RUN.
I THINK PEOPLE DO BLOGGER FOR
INDEPENDENCE.

Here is what was spoken in the previous sample:

THE RULES ARE VERY HARD
FOR ANY THIRD PARTY TO RUN.
I THINK PEOPLE DO HUNGER FOR
INDEPENDENTS.

Consumer survey design

- In section C, errors were highlighted.

- Respondents were asked only if/how the error would affect the respondent's understanding of the caption:

  - No, I understand the caption
  - Yes, it would somewhat affect my understanding
  - Yes, it would greatly affect my understanding
  - Yes, it would completely destroy my understanding

Consumer survey results

- Before seeing the survey's caption error examples, only 11% rated captioning they regularly watch as generally poor overall
- An additional 30% said there were enough significant errors that they sometimes couldn't determine what was spoken
- Only 6% rated captions as "generally accurate"
- In general, the least offensive errors identified were simple substitutions ("mild substitution errors")

Consumer survey results

- The errors in 24 of the sample captions were rated as “severe” by at least half of the respondents.
- The most troublesome errors identified were:
  - Garbling caused by transmission problems
  - Nonsense syllables and words
  - “Major” deletions that impact the meaning of a sentence
  - Gibberish, or collections of letters and syllables that are not legitimate words

Weighted word error rate (WWER)

- Coefficient for each error type
- “Weight” each error type based on survey results
- Align using cc text and clean transcript
- Categorize and calculate

Sample error #2

***** SMOKING DEATH RATES HAVE CONTINUED TO INCREASE**
THE SMOKING DEATH RATE HAS ************ ** INCREASED**
I S S D D S

1 insertion - error type 7 = 0.246
1 singular/plural - type 1 = 0.05
2 wrong tense - type 2 = 2 * 0.057 = 0.114
2 drops (minor) - type 10 = 2 * 0.39 = 0.78

Sample error #2

***** SMOKING DEATH RATES HAVE CONTINUED TO INCREASE**
THE SMOKING DEATH RATE HAS ************ ** INCREASED**
I S S D D S

1 insertion - error type 7 = 0.246
1 singular/plural - type 1 = 0.05
2 wrong tense - type 2 = 2 * 0.057 = 0.114
2 drops (minor) - type 10 = 2 * 0.39 = 0.78

WWER = (.246+.05+.114+.78)/7 = 1.19/7 = 17%

The weighted ranking is more realistic than a simple word error rate.

So…what do we do with all this?

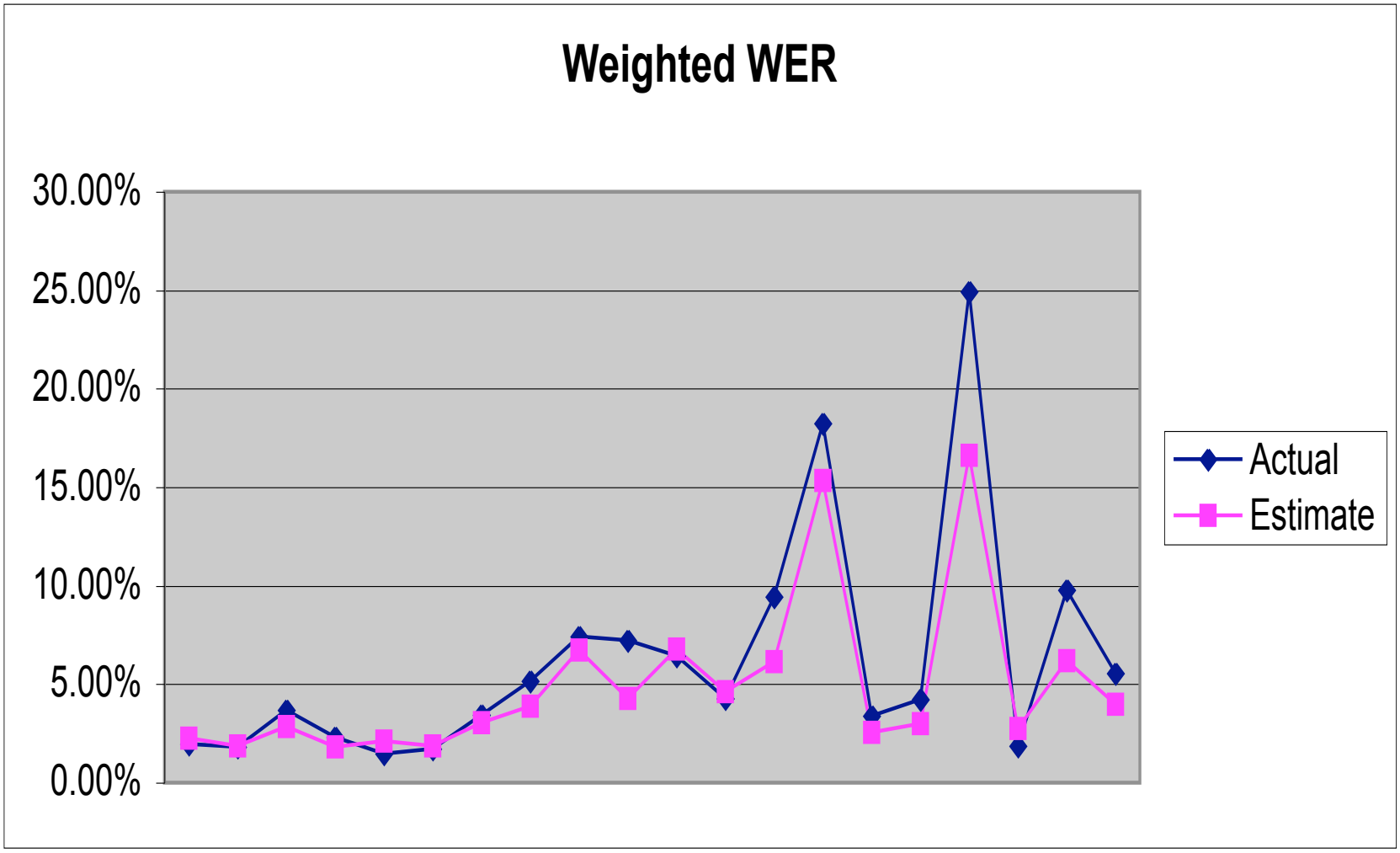
- Still too expensive and time-consuming to manually create a clean transcript for every program analyzed.
- Currently testing whether ASR can be used to estimate the error rate in a caption file.
- When compared to clean transcripts, ASR transcripts are less accurate than captions, but contain different types of errors in different places.

ASR transcript vs. caption file

- Substitutions
  - Usually more substitution in ASR
- Deletions
  - Usually more deletions in captions
- Insertions
  - Few insertions in either

ASR is good at getting a word for each utterance, so total word counts have been accurate.

Preliminary results – 20 programs



Capture station

- DVR for record and playback
- Software (NCAM WordMeter) and hardware to strip caption data and save cc file
- Audio utility to save program audio as mp3 file
- Nuance utility to generate ASR transcript
- Nuance utility to align text & calculate WWER
- Option to add entries to a customized dictionary
- Generate error reports

Year 3 activities

- Compile additional data sets
- Explore further automation of capture station
- Further customize data mining and speech recognition tools
- Beta testing in broadcast environment
- Reconvene technical review panel

- Publish:
  - ✓ Consumer survey summary report
  - ✓ Error ontology
  - *- Research into error capture capabilities of text mining software*
  - *- Reference architecture for prototype evaluator*

The CC Metrics website

http://ncam.wgbh.org/invent_build/analog/caption-accuracy-metrics

Contacts

The WGBH National Center for Accessible Media

Tom Apone
tom_apone@wgbh.org

Marcia Brooks
marcia_brooks@wgbh.org