Caption Accuracy Metrics Project

Caption Viewer Survey: Error Ranking of Real-time Captions in Live Television News Programs

The WGBH National Center for Accessible Media
By Tom Apone, Marcia Brooks and Trisha O'Connell
December 2010

Executive Summary

Real-time captioned television news is a lifeline service for people who are deaf or hard of hearing, providing critical information about their local communities, national events and emergencies. Captioning mandates designed to provide equal access to television have resulted in more accessible programming but a shortage of skilled professionals and the downward pressure on rates by program providers has created the need for a common and automated method of measuring accuracy and quality of real-time captions.

The WGBH National Center for Accessible Media (NCAM) is conducting the Caption Accuracy Metrics project (funded by the U.S. Department of Education, National Institute on Disability and Rehabilitation Research) which is exploring using language-processing tools to develop a prototype automated caption accuracy assessment system for real-time captions in live TV news programming. Such a system could greatly improve the television industry's ability to monitor and maintain the quality of live captioning and ease the current burden on caption viewers to document and advocate for comprehensible captions to ensure they have equal access to important national and local information. Ideally, this system will be able to differentiate between stenocaption errors and technical errors; identify types and frequency of stenocaption errors (e.g., mistakes, word deletions or substitutions); quantify the display delay between the spoken word and the associated caption; and indicate whether words and phrases that were spoken are missing entirely. An additional challenge for the system will be identifying errors that radically impact the comprehensibility of news programming.

In spring 2010, NCAM conducted a national Web survey to query television news caption viewers about the types of caption errors that impact their ability to understand a live television news program. Survey results are contributing to definition of error types and criteria for weighting and ranking error types within the prototype automated caption accuracy assessment system.

Over 350 caption viewers from across the U.S. completed the survey. The majority of respondents self-identified as deaf or late-deafened; less than a third indicated they were hard-of-hearing. The survey presented 41 examples drawn from a wide range of major national broadcast and cable television live news programs. These 41 examples represented 17 sub-categories of common caption error types identified by the project team and advisors. Errors in 24 of the 41 examples were rated as severe by at least half the respondents. Severe errors included: garbling caused by transmission problems, nonsense syllables and words caused by stenocaptioner error, and major deletions that impact the meaning of a sentence. The least problematic errors were simple substitutions (such as the wrong tense) and errors in punctuation.

Survey Design

The survey was designed in consultation with project advisors Judith Harkins, Ph.D., professor in Gallaudet University's Department of Communication Studies and founder of the Technology Access Program, and James J. DeCaro, Ph.D., Professor and Dean Emeritus, Interim President, National Technical Institute of the Deaf at the Rochester Institute of Technology. The initial design of our draft survey presented captioned video clips to respondents. Beta tests with this draft surfaced potential problems using video clips. Some respondents felt they were being tested on their caption-reading skills and many chose to watch the clips multiple times before answering each question. The goal was to ask each respondent's thoughtful opinion about the impact of different types of errors, not to test caption-reading skills. The goal was also to keep the average time to complete the survey between 20 and 30 minutes. Consequently, in the final survey, examples were presented to respondents in a still video frame or as plain text, rather than video clips. Viewing caption errors this way allowed respondents to focus on their ability to decode the caption error when answering each question.

The final survey consisted of three sections:

Section A collected demographic information about the respondents (e.g. age, type of sensory disability if any); their television viewing habits; and their overall impressions of caption accuracy (in general, and specifically in television news programs).

Section B presented 15 caption examples from actual network broadcast and cable television newscasts, which were presented as still frames with two questions asked about each example. For each caption example in sections 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)

If a respondent 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

Example of a still-frame caption sample used in the survey:

Section C consisted of 26 more caption examples presented as text only, along with the correct text of what was actually spoken. Respondents were asked to compare the two and rate how the error would affect their understanding of the content. Since the errors were highlighted, a single question asked if 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

Example of a text caption sample used in the survey:

Captions as they appeared onscreen:

IT WAS A VERY SMART PLAY**.**

Here is what was spoken in the previous sample:

IT WAS **ACTUALLY** A VERY SMART PLAY.

Survey Recruitment

NCAM distributed the survey invitation to disability-focused listservs and blogs as well as NCAM's own extensive list of individual and organizational contacts. In addition, many national and regional consumer advocacy groups re-distributed the survey invitation to their lists of constituents. The invitation specifically solicited only people who use captions when viewing television news programs.

Survey Respondents

Over 900 people visited the Web site during the three-week period the survey was open (early April 2010). Nearly half of those visitors (422) abandoned the site without taking the survey. Since the introductory page of the survey indicated the survey was specifically for those who regularly use captions to watch live newscasts, many visitors who did not fit the criteria stopped.

352 people completed the survey. An additional 144 visitors responded to some of the questions but did not complete the entire survey. Respondents were able to fill out the survey in increments and return to the survey for further entry.

The majority of respondents (50%) self-identified as deaf. An additional 12% selected late-deafened as their hearing status and 29% indicated that they were hard-of-hearing. Just over 30% of respondents indicated they also had mild or moderate vision loss.

Nine percent of respondents identified themselves as hearing; of these 34 hearing respondents, nearly half indicated they regularly watch captioned TV with companions who require closed captions. The remaining hearing respondents indicated a range of reasons for regularly watching captioned TV, from “watching TV quietly to avoid disturbing others in my household” to “watching captions in public spaces where the TV volume is not loud enough or is not turned on.”

Respondents by Self-Identification:

	# of People	Percent
Deaf	175	50%
Late Deafened	42	12%
Hard of Hearing	101	29%
Hearing	34	9%

The majority of respondents (62%) were between 30 and 60 years old. An additional 34% of respondents were over 60. Only 4% were under 20 years of age.

Seventy-four percent of all respondents said they watch one or more television newscasts every day.

Caption Error Types Represented in the Survey

Stenocaption error assessment is traditionally broken down into three main categories: Substitutions, Deletions, and Insertions. For real-time captioning, deletions (also known as drops), where spoken words are omitted from the text, are usually the most common error. If deletions are minor (an aside like “you know”), they may not significantly impact meaning. In some cases, though, captioners drop larger passages and delete important information.

Substitutions are also common and represent a wide range of errors. If the substituted word is close to the original (a homophone or slight misspelling), it may not have a major impact. Because stenography is based on phonetics, many substitutions have some phonetic similarity to the correct spoken word. More severe substitutions include a wrong word that changes the meaning of a sentence or a collection of nonsense syllables or letters that is completely unintelligible.

Insertions are rare, most often occurring in conjunction with another error.

These three main error categories were parsed into 17 error types, based on extensive analysis of many sample texts. We also drew on the National Court Reporters Association (NCRA) and its grading system for identifying errors. The 17 error types were presented in a total of 41 sample captions in this survey. In gathering real world examples of caption errors from national broadcast and television network news, we attempted to present only those with one error per example.

Caption Error Types:

		Substitution	Deletion	Insertion
1	Substitute singular/plural	Yes		
2	Substitute wrong tense	Yes		
3	Sub pronoun (nominal) for name	Yes		
4	Substitute punctuation	Yes		
5	Split compound word, contraction (OK)	Yes		Yes
6	Two words from one (one wrong)	Yes		Yes
7	Duplicate word or insertion			Yes
8	Word order		Yes	Yes
9	Correction by steno			Yes
10	Dropped word - 1 or 2		Yes	
11	Dropped word(s) - 3+		Yes	
12	Homophone	Yes		
13	Substitute wrong word	Yes		
14	Not a valid word	Yes		
15	Random letters (gibberish)	Yes		
16	Word boundary error	Yes		
17	Transmission errors/garbling	Yes		

Error Ontology – Substitutions

The first group of errors (errors 1-4) contains mild substitutions that are commonly seen in real-time captions. Examples include simple grammatical errors and subjective punctuation decisions such as when to use a question mark (the latter tend to be style choices by a caption stenographer and/or transcriber capturing the transcript's actual spoken words).

Error Ontology – Mild Substitutions:

	Error Type	Example (caption/actual)
1	Substitute singular/plural	man/men
2	Substitute wrong tense	run/ran
3	Sub pronoun (nominal) for name	this man/Proper Name
4	Punctuation differences	(period instead of a question mark)

More significant substitutions (errors 12-17) can range from simple homonyms (e.g., sail/sale) to more subtle changes (e.g., feature/future) to nonsense words and phrases (e.g., photostat us quo/for the status quo) and finally to complete gibberish. Nonsense words and phrases can be very difficult for viewers because the words in the sentence may be valid, they just don't make sense. A viewer struggles to interpret, ultimately cannot and in the meantime, has fallen behind in reading the captions and loses more information. Because comprehensibility is more difficult to define for these complex errors, the survey was more heavily weighted with samples from this group.

Error Ontology – Severe Substitutions:

	Error Type	Example (caption/actual)
12	Nearly same sound but wrong word(s)/homophone	sale/sail or work/werk
13	Substitute wrong word	Blogger/hunger
14	Phonetic similarities, not valid words	human milating/humiliating
15	Garbled syllables, not words	igbavboa
16	Word boundary error (also "stacking error")	paying backpack Stan/paying back Pakistan
17	Transmission: paired letter drop, white boxes, garbling	GM sto/GM stock

Error Ontology – Insertions

The second group of errors (5-9) contains insertions. These usually occur in conjunction with another type of error and can often be decoded in context.

Error Ontology – Insertions:

	Error Type	Example (caption/actual)
5	Split of compound word, contraction (OK)	foot note/footnote did not/didn't
6	Two words from one (one wrong)	might yes/mighty
7	Duplicate word or minor insertion	criticism criticism
8	Word order/transposition	would I/I would
9	Correction by steno	disznlts—dissidents

Error Ontology – Deletions

The third group of errors contains deletions. We categorized these either as minor or significant deletions. There are examples where a single word is dropped and that omission changes the entire meaning of a sentence. However, stenocaptioners often drop minor asides such as “well” or “you know” or drop additional adjectives and modifiers that are unnecessary or redundant. Omitting larger phrases or entire sentences can have a more significant impact on the correct transmission of meaning.

Error Ontology – Deletions:

	Error Type	Example (caption/actual)
10	Dropped word(s): 1-2 (minor, aside)	“you know”
11	Dropped word(s): 3 (or significant)	“figure out what the best options are going to be”

Survey Results

Pre-Survey Opinions

At the beginning of the survey in Section A, before respondents were presented with error samples, respondents were asked to indicate their overall opinion of caption quality. Respondents were asked to select one of a number of statements that best reflected their general opinion about caption errors in live newscasts. Over 50% indicated that many real-time caption errors are minor but 42% indicated that caption errors negatively impacted their ability to understand what was spoken. Only 6% of respondents felt that real-time captions were generally accurate.

Opinions Prior to Viewing Survey Samples:

	# of People	Percent
I think the captions are generally accurate	21	6%
I think there are some minor errors	75	21%
I think there are a lot of minor errors, but I can still determine what was spoken.	107	30%
I think there are a few significant caption errors that change the meaning of the spoken word(s) and I sometimes can't determine what was spoken	108	31%
I think there are many significant errors and I often cannot determine what was spoken	40	11%

Overall, these opinions about caption errors in live newscasts showed a similar spread across all four populations with one noteworthy difference. Deaf viewers were more likely to indicate that they felt captions generally contain some minor errors and less likely than any other population including hearing viewers to agree that there are many significant caption errors in live newscasts that affect their ability to determine what is being spoken.

Opinions Prior to Survey by Population:

	Deaf	Late Deaf	Hard of Hearing	Hearing
I think the captions are generally accurate	7%	0%	4%	12%
I think there are some minor errors	26%	17%	19%	12%
I think there are a lot of minor errors, but I can still determine what was spoken.	28%	36%	33%	30%
I think there are a few significant caption errors that change the meaning of the spoken word(s) and I sometimes can't determine what was spoken	31%	31%	31%	27%
I think there are many significant errors and I often cannot determine what was spoken	7%	17%	14%	18%

It may be that the residual hearing of late-deafened or hard-of-hearing respondents provides clues as to the degree of missed content, leading to higher estimates of the number of significant errors. It also may be that deaf viewers develop caption-reading skills that decode or fill in partial content. Late-deafened, hard-of-hearing and hearing caption viewers were at least twice as likely to indicate they saw many significant errors and had difficulty determining what was spoken.

Rating Caption Errors

In general, the survey results aligned with the ranking of error types (mild to severe) as identified in the draft error type ontology.

We defined a “severe error” as one that more than 50% of respondents identified as greatly impacting or completely destroying their understanding of the sentence. The errors in 24 of the sample captions were rated as severe by this measure.

The most troublesome errors identified were garbling caused by transmission problems, nonsense syllables and words, and “major” deletions that impact the meaning of a sentence.

Below are the seven error types with the worst ratings (the percentage of respondents rating the caption as “greatly affecting” or “completely destroying my understanding”):

Error types with the worst ratings:

Error type	Description	% of respondents
17	Transmission error, dropped letters	84%
16	Word boundary error	65%
15	Garbled syllables, not words	65%
14	Phonetic similarities (not words)	59%
11	Dropped words (significant, 3+)	59%
13	Substitute wrong word	55%
3	Substitute pronoun for proper name	53%

The ten remaining error types did not meet the 50% “severity” threshold, though three more (error types 6, 8, and 9) rated in the mid-40s percentile. The full rankings are available below.

Survey Results by Caption Error Types

Error types 1-4 (mild substitutions): In general, less than 10% of respondents considered these errors to have a major impact. Error type 3, however (substituting a pronoun or nominal for a proper name), was the exception – it was rated a significant error by more than half of the respondents. This is not a surprising finding since we know anecdotally that many caption watchers are frustrated by this practice. It is usually the result of insufficient prep time and/or insufficient advance information from the program producer. Quantifying this substitution rate can help indicate when preparation time and advance information is lacking.

Error types 5-9 (insertions): Errors in this group were generally rated as moderate problems. None reach the 50% threshold to be classified as severe. As noted earlier, these errors contain an insertion but also frequently include a substitution as part of the error. For example, when the captioner wrote “might yes” in place of the spoken word “mighty,” might is considered a substitution for mighty and yes is considered an insertion or additional word.

Error types 10-11 (deletions): Errors become more subjective and contextual for deletions and more severe substitutions. For deletions, the particular words that are dropped have a dramatic impact on whether the sentence is understandable or accurate. Certainly, the meaning of a sentence can be completely changed by the deletion of a single word. In most cases, however, small deletions (drops of a word or two) are asides or additional modifiers that do not alter meaning significantly. In one of our survey examples, a speaker in a newscast said, “It was actually a very smart play.” The captioner omitted the word actually and the caption read, “It was a very smart play.” In this case, respondents graded this error as mild.

Deleting or dropping more than three (contiguous) words in a sentence (error type 11) was rated the most significant type of error in this group. While it is true that in some cases an entire phrase or sentence may be dropped and captions might still accurately reflect the “gist” of a sentence or passage, when large amounts of text are dropped some meaning is inevitably lost.

There are even times when a paraphrased caption is more readable than the original spoken word, but extensive editing and paraphrasing is not a good practice and is usually viewed negatively by caption watchers.

Error types 12-17 (severe substitutions): This group of errors represents substitutions that were rated as most egregious. These typically have a negative, even misleading, impact on comprehension. This group of errors presented the most difficulty to viewers.

We used four different types of substitutions in the survey to test whether respondents would distinguish these differences. Error types 12-15 represent substitutions with an increasing level of severity.

Error type 12, a homophone, can typically be understood by most viewers. (Some viewers may not even identify these as errors or may, in fact, misuse things like “they’re” vs. “their” in everyday communication.) Similarly, a slight misspelling (“werk” in place of “work”) is usually understandable in context.

Error type 13 becomes more problematic for viewers. Here, an actual word is substituted but it is not close enough to the spoken word to convey the correct meaning. In extreme cases it may even make some sense but mean something different than what was spoken.

Error types 14 and 15 are the most extreme cases of substitutions. Error type 14 may include part of words, nonsense words or things that look like words, as in Lewis Carroll’s “The Jabberwocky”. Errors of type 15 have no meaning and usually were no more than random letters grouped together.

Error type 16 represents some unique situations where compound words are broken and recombined across word boundaries. Typically, the new words that have been created take on different meanings from the original spoken word and are very difficult for viewers to interpret.

There was not a significant difference in how respondents ranked errors 14, 15 and 16.

Error type 17 represents a special category -- garbled or corrupted caption data caused by transmission problems. These can look similar to a real-time captioning error or a misspelling. Savvy caption watchers are familiar with the garbled text and white boxes that frequently appear in these cases and anyone who has watched more than a few minutes of badly garbled captions will soon give up. Respondents rated these errors the most troubling by a large margin.

Full Survey Results

We considered responses that “greatly affect my understanding” or “completely destroy my understanding” to indicate a severe or unacceptable error.

Here is the ranking of error types based on the number of respondents who thought the sample caption would “greatly affect” or “completely destroying my understanding” of the content:

Error types ranked by severity:

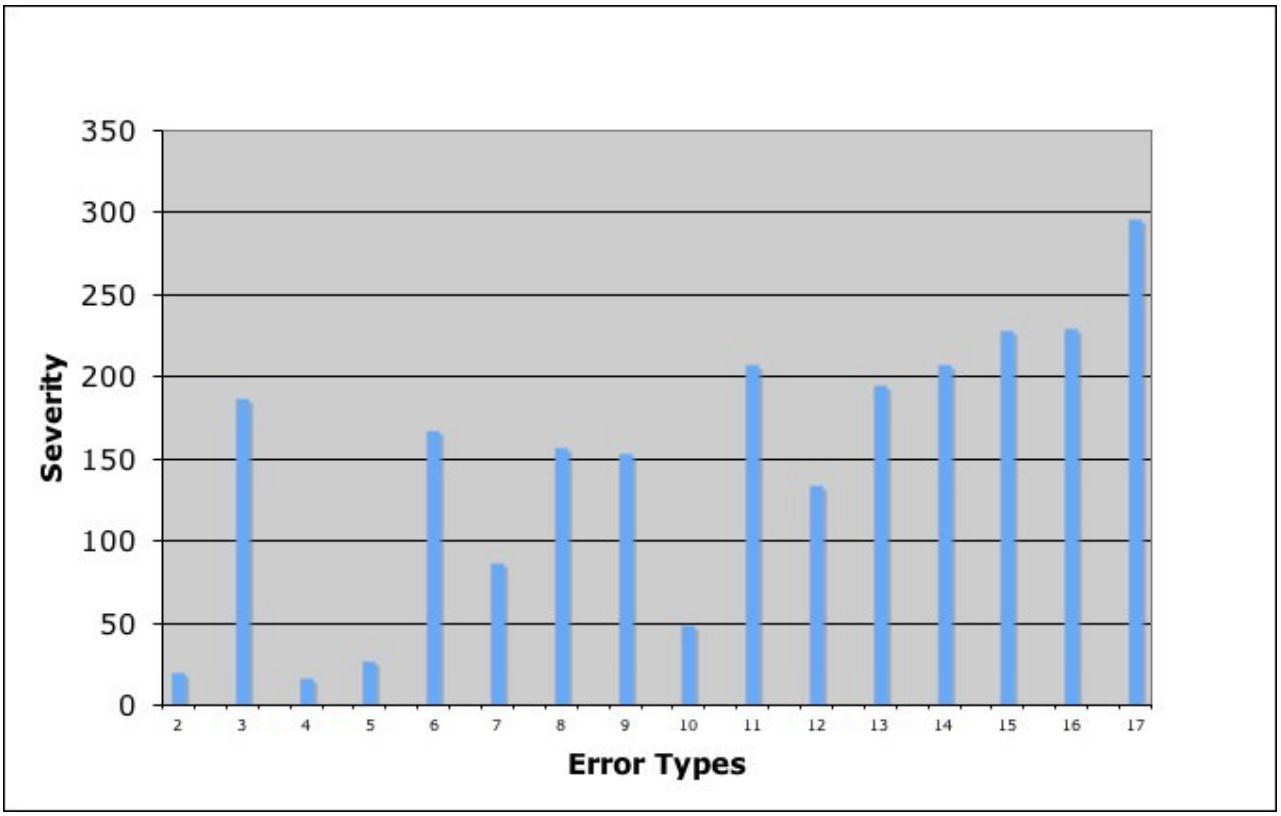
Error Type	Description	%
17	Transmission errors	84.3%
16	Word Boundary (also stacking error)	65.2%
15	Garbled syllables, not words	65.0%
14	Similar sounds and syllables (steno)	59.1%
11	Dropped Words: Major (3+)	58.9%
13	Substitute Wrong Word	55.4%
3	Sub pronoun (or nominal) for Proper Name	53.3%
6	Two words from one (one word wrong)	47.4%
8	Word Order/Transposition	44.7%
9	Correction by Steno	43.7%
12	Homophone	37.9%
7	Duplicate Word/ minor insertion	24.5%
10	Dropped Words: Minor, 1-2 (aside)	13.7%
5	Split Compound Word, Contraction	7.7%
2	Wrong Tense	5.7%
4	Punctuation	4.6%
1	Singular/Plural	N/A

In general, the ranking of error severity followed the pattern one might expect. Punctuation, splitting compound words, and tense differences were not judged to be major problems. Dropping significant amounts of text and substitutions that were unintelligible or changed meaning were judged to be significant problems.

Error type 3 (substituting a pronoun or nominal for a proper name) was also ranked severely. These substitutions allow for the construction of a complete sentence in cases where the stenocaptioner does not know the spelling of the proper name. However, caption watchers can be deprived of important information when a person’s name isn’t made available to them in the caption text. So, while the caption itself may have been readable, our respondents indicated that it had a negative impact on their understanding.

Error type 9 (correction by the captioner) was also judged to be a significant problem. While a good stenocaptioner can make corrections “on the fly” these still typically come after the error is displayed and this correction process may be jarring or difficult for viewers to follow. However, this category still ranked significantly better than an error that was not corrected.

Error Types v. Error Severity



“Severity” is the number of respondents rating an error as “greatly affecting” or “completely destroying” understanding.

Conclusions

There is a wide range of error types in real time captioning and they are not all equal in their impact to caption viewers. Treating all substitution and deletion errors the same does not provide a true picture of caption accuracy. These results provide valuable data about how to rank the severity of the 17 types of errors evaluated through this survey. The least offensive errors were judged to be simple “substitutions” like the wrong tense and punctuation; however, substituting pronouns and/or nominals for proper names were also judged to significantly impact viewers’ understanding.

The most troublesome errors were judged to be garbling (a form of substitution) caused by transmission problems, nonsense syllables and words, and “major” substitutions and deletions that impact the meaning of a sentence. Substitutions present the most varied and difficult type of error to parse into more detailed categories. We identified 12 different kinds of substitution errors, from simple punctuation changes to more severe word substitution errors. The four errors rated as most severe were from the main category of substitution errors, followed in fifth place by a major deletion of three or more words.

Many serious errors stem from transmission and equipment problems in the broadcast chain, exacerbated by the transition to digital television. New data streams, new and untested equipment, and multiple signal transcoding have all contributed to technical difficulties, some of which are currently being addressed by the Federal Communications Commission (FCC).

The final phase of our current project involves developing automated measuring tools for evaluating caption quality. We are developing a Weighted Word Error Rate that will rank caption text based on the relative weightings from the survey data. Initially, this measurement is being accomplished using a manually created clean program transcript that is aligned with the caption text for comparison. We are also testing whether speech recognition systems can be used to estimate a Weighted Word Error Rate without the need for a clean transcript.

Going forward, the challenge is to design a software system that can automatically identify these errors and rank each according to its impact on viewers’ comprehension. For most of the errors in the list, this will be a readily achievable task. Language processing tools are increasingly skilled at identifying garbled words, pronoun substitution or incorrect syntax. Pronoun substitution is an example of an error relatively easy to identify but one that warrants a serious error in ranking.

The more difficult task is to fine tune a system that can differentiate between phonemes and more severe substitutions or flag whether deletion of a one-word modifier is minor or radically changes the meaning of the sentence. Similarly, on-the-fly corrections made by stenocaptioners will be relatively easy to identify but may require additional analysis to determine the impact on comprehension.

The final, proposed calculation for word error rate will be included in the research report we will publish upon conclusion of the project. This report will address the error capture capabilities of text mining software agents and the customized rules and classifications that will be derived from the stenocaption ontology. Reference the Caption Accuracy Metrics web site for a full list of project deliverables.