RT-S: SURFACE RICH TRANSCRIPTION SCORING, METHODOLOGY, AND INITIAL RESULTS

Matthew Snover†, Richard Schwartz‡, Bonnie Dorr†, John Makhoul‡

†Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740

‡BBN
9861 Broken Land Parkway
Columbia, MD 21046

ABSTRACT

In this paper we present a methodoly for the scoring of punctuation annotated texts, as well as a preliminary system to perform the task. We modify SCLITE's scoring methodology to support scoring of punctuation. Using this methodology, we show that the error rate of an initial automatic system is comparable to annotator inconsistency. However, the use of multiple references allows us to differentiate between human inconsistencies and system errors.

1. INTRODUCTION

One of the goals of the EARS[1] (Effective, Affordable, Reusable Speech-to-Text) program is the development of systems for the automatic augmentation of speech transcripts with metadata[2]. Previous approaches to this [3, 4, 5, 6] have focused on the underlying metadata, such as edit events and interruption point events. These approaches do not, directly, generate a more human readable transcript.

The goal of Surface Rich Transcription (RT-S), formerly known as RT-A, is to produce a complete and readable transcript of speech using normal writing conventions. The transcript is complete, in that it contains everything that was said, including disfluencies and significant non-speech events. The transcript should be intelligible despite the presence of these disfluencies, and this information should be conveyed using normal writing conventions. While previous studies of the human readability of speech transcripts[7] have not shown a large quantitative benefit to the addition of metadata to speech transcripts, it does seem clear that there is a subjective preference for punctuated, capitalized transcripts.

1700 hours of Fisher training data has been annotated by WordWave according to the RT-S guidelines, giving a large base of training data for automatic systems designed to augment transcripts with RT-S information. In addition, the Ears Evaluation 2003 Fisher data has been annotated by WordWave using 5 sets of annotators, allowing us to conduct an in-depth inter-annotator consistency study, as well as explore the possibly of using multiple references for scoring.

This paper first describes the RT-S task, and the recent modifications to the Punctuation Style Guides. We then describe a modification to the standard SCLITE alignment tool that allows us to score to the punctuation of RT-S transcripts. Using this scoring method, we evaluate inter-annotator consistency, and then evaluate a preliminary automatic punctuation identification system. The scoring method proposed does not seem adequate for the task, and so we describe a system of scoring using multiple annotators, which enables us to more clearly discriminate the acceptable differences in human annotations from automatic system errors. Finally we show the results of our system when used with automatic speech recognition (ASR) transcripts.

2. TASK

Automatic Speech Recognition systems typically output a sequence of words with no punctuation, speaker turns, or capitalization. RT-S aims to add all three of these properties to automatic transcripts in the aim of improving readability. Because we focus solely on conversational telephone speech (CTS), speaker identification is straight-forward, since each speaker is observed on a separate audio channel. One speaker is labeled as the left channel (L) and the other as the right channel (R). The left and right channels are allocated on the basis of which side of stereo output the sounds were produced during transcription.

The current task of the RT-S system is to recognize five types of punctuations: commas, periods, exclamation points, question marks, and discontinuities, which are marked using a double hyphen (--). In addition continuations, represented by ellipses (...) are also annotated, but these are just formatting purposes and are not scored. While speaker turns and capitalization are also part of the RT-S task, we are not examining them in this work.

Figure 1 contains a sample output from an ASR system, and figure 2 contains a sample of that same transcript marked up in RT-S format.

A: um so I'm thinking for the second part which i i guess you didn't hear i'm thinking the if i had to make up a holiday we'd combine halloween and christmas and it's like a nightmare before christmas have you seen that movie
B: mhm [LAUGH] yes

Fig. 1. Sample ASR Output

L: Um, so I'm thinking for the second part...
R: Mhm.
L: ... which I -- I guess you didn't hear. I'm thinking the -- if I had to make up a holiday we'd combine Halloween and Christmas...
R: [LAUGH]
L: ... and it's like a Nightmare before Christmas -- have you seen that movie?
R: Yes.

Fig. 2. Sample RT-S Transcript

2.1. Punctuation Style Guidelines

A few modifications have been made to the RT-S Punctuation Style Guide. The 1700 hours of Fisher data, and the EARS Evaluation 2003 data were both annotated according to the previous guidelines, but the to the extent possible, these changes have been automatically added to these corpora.

Filled pauses, such as "um" and "uh" and interjection fillers, such as "you know" are surrounded by commas, if there is no other punctuation present. For example: "The guy went, uh, to his friend's house, you know."

Discontinuities are now marked with a "--". These occur at the beginning of edits, such as repetitions and restart, as in the case of "I'm going -- I went there." and "Wait, let me just ask you this -- hold on, what was it? Oh yes, did you -- were you ever in Paris in the springtime?" Discontinuities are also marked with the speaker is cut off and does not continue such as in the example in Figure 3. Discontinuity markers are not used immediately after fragments.

L: What I was about to say --
R: Do you know what time it is?
L: Yes, it's ten past five.

Fig. 3. Discontinuity Example

Continuations are marked with "...". These represent over-speaking, when two speakers speak at the same time, but the interrupted speaker continues what they were previously saying. An example of a continuation is shown in Figure 4.

L: What I was about to say...
R: Wait a second!
L: ... was that I never want to see you again.

Fig. 4. Continuation Example

3. PUNCTUATION SCORING

Punctuation accuracy is scored by aligning the hypothesis transcript to a reference transcript using a modified version of SCLITE, with all punctuation being separated from the words and treated as separate tokens. The output of SCLITE contains both punctuation errors and word errors, so another program is used to separate the two error types.

The weights used for the dynamic alignment within SCLITE have been modified, so that the Punctuation Error Rate (PER) is minimized under the constraint of minimum Word Error Rate (WER). The cost of aligning punctuation to words has been raised so that two cannot be aligned to each other, and the cost of making punctuation errors has been made 100 times less that of making words errors. We built a separate tool that extracts and counts the word errors and punctuation errors from SCLITE's alignment.

For all of the scoring reported in this paper, no global language model (GLM) was used. This causes a slightly higher word error, and the possibility of a slightly lower punctuation error rate, than one would observe if using a GLM file.

4. INTER-ANNOTATOR CONSISTENCY

We compared the five annotators for the EARS Eval 2003 Fisher Data against each other. Each annotator was scored against each of the other annotators. The average WER was 10.5%. The full matrix of PER scores is shown in Table 1. The average PER was 46.8% - a very high disagreement rate. No real outliers were observed.

A break down of punctuation by type as well as by insertion/deletion/substitution rate is shown in Table 2. All of the values shown in that table are averaged over the 20

	Ref. 1	Ref. 2	Ref. 3	Ref 4	Ref. 5
Hyp. 1	0.0	43.8	46.9	44.7	46.7
Hyp. 2	43.3	0.0	52.9	45.1	48.0
Hyp. 3	39.9	45.5	0.0	43.5	47.8
Hyp. 4	46.9	47.8	53.7	0.0	56.3
Hyp. 5	42.0	43.9	50.4	48.2	0.0

Table 1. Inter-Annotator PER

combinations references and hypothesis (scoring an annotator against themselves was not used in the average). The number of substitution errors is less than either the insertion or deletion errors for all punctuation, indicating that annotators have trouble agreeing on the location of punctuation, rather than the type of punctuation for a location. This is not true for periods, question marks, and exclamation points though, all of which have higher substitution error rates than insertion or deletion error rates.

Punct. Type	Occ	%Sub	%Ins	%Del	%Err
Words	33524	5.3	2.6	2.6	10.5
All Punctuation	8812	12.7	17.4	16.7	46.8
Commas	4213	10.7	24.5	23.0	58.1
Discontinuities	1778	14.7	18.8	17.4	50.8
Periods	2485	13.8	7.1	7.0	27.9
Question Marks	323	15.2	5.0	4.9	25.1
Exclamation Pts	12	72.6	22.0	6.6	96.2

Table 2. Inter-Annotator Error by Type

Table 3 is a confusion table showing the substitutions for each punctuation type. This table shows that periods and commas are often confused by annotators and that discontinuities and commas are also often confused.

	Null	Comma	Discont	Period	Ques	Excl Pt
Null		990	316	175	16	1
Comma	990	2779	188	238	17	2
Discont	316	188	1194	73	6	1
Period	175	238	73	1964	27	8
Ques	16	16	6	27	258	9
Excl Pt	1	2	1	8	0	1

Table 3. Inter-Annotator Error Punctuation Confusion

5. A PRELIMINARY PUNCTUATION ID SYSTEM

Our preliminary punctuation ID system takes as input a word transcription and outputs a punctuated sequence of words. Speaker turns are given to the automatic system from the word sequence, if available, and are otherwise, in the case ASR input, estimated by breaking at long silences.

We treat the punctuation identification task as a part-of-speech tagging task with punctuation instead of parts of speech. Each word in the training data is labeled with the punctuation that comes immediately after it, or with the NULL label if there is no punctuation. We used a maximum entropy part of speech tagger[8] trained on 2 million words (about 200 hours of speech) from 1700 hours of Fisher data. The maximum entropy tagger primarily uses the immediately surrounding words as features for identifying the correct tag or punctuation to be applied. The error rate on the training data was a suprisingly high 43.3% PER. This is mostly due to the amount of variation in punctuation annotation, and the limited feature set that the model used.

For each annotator, we ran the Punctuation ID system on the words from the transcript of that annotator and scored the results against all of the other annotators. The word error rate for these experiments was exactly the same as in the inter-annotator experiment, as none of the words had changed, and the word error rate is always minimized. A full matrix of the PER results of this experiment is shown in Table 4. This gives us a table of PER results for comparing each annotator's words with automatic punctuation against annotator's words and reference punctuation. When comparing against other annotators, so that the words are different, the average PER of the system was 56.3 (these are the non diagonal values). When comparing against the same annotator, so that the WER is 0.0, the average PER of the system was 48.6. Only the result with different reference words is directly comparable against the inter-annotator results. The results with the same reference words more accurately reflect the task that we are interested in measuring, namely how accurate is the punctuation ID system, given that the words are correct.

Sample output from the Punctuation ID System is shown in figure 5. There are five errors in the example, out of a total of nine punctuation marks. Three of the errors are deletions, and the other two are substitutions.

	Ref. 1	Ref. 2	Ref. 3	Ref 4	Ref. 5
Hyp. 1	44.7	52.7	51.7	52.9	52.4
Hyp. 2	55.3	46.6	55.4	55.2	55.8
Hyp. 3	54.5	55.9	48.8	55.9	56.2
Hyp. 4	60.3	60.7	60.0	52.4	61.7
Hyp. 5	56.5	57.4	57.0	58.6	50.3

Table 4. Punctuation ID System PER

The results of the automatic system evaluation by punctuation type, averaged over the various trials, is shown in Table 5. While the automatic system performed worse than the humans overall, it did have fewer errors than humans for exclamation points and commas. Question marks proved especially hard for the automatic system to identify, probably due to limitations in the maximum entropy model.

A comparison of the punctuation substitution tables for the Automatic System with different reference words in Table 6 and the Automatic System with the same reference

Punct Type	Inter-Ann Diff Ref	Auto Sys Diff Ref	Auto Sys Same Ref
All Punct	46.8	56.3	48.6
Commas	58.1	53.4	45.5
Discont	50.8	83.3	76.5
Periods	27.9	39.1	29.7
Question	25.1	79.2	77.0
Excl Pts	96.2	83.3	80.0

Table 5. Inter-Annotator Error by Type

words in Table 7 show no real differences in the patterns of substitutions. The pattern of substitutions, actually, are very similar for the automatic system and the human annotators. One difference though, that the Automatic System deletes more periods that it substitutes, resulting in longer sentences in the output.

	Null	Comma	Discont	Period	Ques	Excl Pt
Null		656	104	180	8	0
Comma	1356	2632	38	182	4	0
Discont	1022	241	403	105	7	0
Period	432	289	53	1692	20	1
Ques	62	28	15	143	75	0
Excl Pt	1	3	0	9	0	0

Table 6. Punctuation Confusion for Automatic System with Different Reference Words

	Null	Comma	Discont	Period	Ques	Excl Pt
Null		529	58	63	3	0
Comma	1197	2830	21	163	2	0
Discont	982	200	478	111	6	0
Period	346	265	2	1808	24	1
Ques	53	23	14	155	78	1
Excl Pt	1	2	0	10	0	0

Table 7. Punctuation Confusion for Automatic System with the Same Reference Words

```
REF:  Um , so I'm thinking for the second part
      which I – I guess you didn't hear . I'm thinking
      the – if I had to make up a holiday we'd combine
      Halloween and Christmas and it's like a Nightmare
      before Christmas – have you seen that movie ?
      Yeah . Um , something like that would be pretty cool .
HYP:  Um , so I'm thinking for the second part
      which I – I guess you didn't hear * I'm thinking
      the ** if I had to make up a holiday we'd combine
      Halloween and Christmas and it's like a Nightmare
      before Christmas ** have you seen that movie .
      Yeah , Um , something like that would be pretty cool .
```

Fig. 5. Sample Automatic Punctuation System Output

6. MULTIPLE REFERENCES

The Inter-Annotator PER and Automatic System with different reference words are close enough that one could wonder if the Automatic System is almost as good as a human. A subjective look at the results of the automatic system show that while it appears better than no punctuation it does make a number of errors that a human would never make, such as not placing any punctuation in long word sequences, and missing a large number of question marks. To address this issue we have developed a method for scoring with multiple references. If a token in the hypothesis agrees with any of the references, then it is marked as correct. The denominator, when calculating PER and WER in this way, includes tokens that were deletable due to the multiple references.

We generate a combined reference by successively aligning each reference to the other references using the modified version of SCLITE. We combine the aligned reference into a single reference with multiple choices for some punctuation and words, using SCLITE's built-in tools for alternations in references. We can then score a hypothesis in the same method as before: by aligning it with SCLITE, and then using our tool to extract the word and punctuation errors. An example of multiple references is shown in Figure 6, with multiple values for a token being shown in curly braces. The "@" symbol represents a deletable token.

```
REF:  i do { . / , } it's halloween . { uh / ah } , i think
      it's because i get to be something that i'm
      not { , / @ } and so do other people . i really
      like ** ** to see other people on halloween
HYP:  i do , it's halloween . uh , i think
      because i get to be something that i'm
      not and so do other people . i really
      like TO - - to see other people on halloween .
```

Fig. 6. Multiple Reference Example

The results of evaluating both inter-annotator agreement and the automatic system using multiple references is shown in Table 8. We compared each annotator or system output against the combination of the other 4 annotators. Using multiple references there was about a 36% absolute error reduction in both cases. The ratio of automatic system error to inter-annotator error is higher when using the 4 combined annotators as reference. Using multiple annotators for reference, the ratio of inter-annotator to automatic system PER went from 1.2 to 1.7, giving us a better method of discriminating the poorer automatic system output from the inter-annotator agreement.

We cannot evaluate the automatic system with the same reference words for the 4 combined annotator reference, because there is no single sequence of reference words, upon which the system can be run. This could be accomplished by training a new system on transcripts with word alternations, however this would require far more training data than is currently available.

	Inter-Ann Diff Ref	Auto. Sys Diff Ref	Auto. Sys Same Ref
1 Annotator	46.8 10.5 WER	56.3 10.5 WER	48.6 0.0 WER
4 Annotators Combined	11.4 3.8 WER	19.1 3.8 WER	–

Table 8. Multiple Reference Results

7. AUTOMATIC SPEECH RECOGNITION RESULTS

The ultimate aim of RT-S is to use an automatic punctuation identification system to enhance ASR transcripts. We added punctuation using our automatic system to ASR output from BBN's Byblos Speech Recognizer. We then scored these transcripts using both 1 and 4 annotators. The results are shown in Table 9.

	Inter-Ann Diff Ref	Auto. Sys. Diff Ref	Auto. Sys. Same Ref	Auto. Sys. ASR Words
1 Annotator	46.8 10.5 WER	56.3 10.5 WER	48.6 0.0 WER	73.3 25.4 WER
PER - WER	31.3	45.8	48.6	47.9
4 Annotators Combined	11.4 3.8 WER	19.1 3.8 WER	–	32.3 18.8 WER
PER - WER	7.6	15.3		13.5

Table 9. Automatic System ASR Results

With the four combined annotators, the automatic system with ASR is 1.7 times worse than with reference words, and 2.8 times worse than the Inter-Annotator agreement. It is clear from this, and our previous results, that different words cause different punctuation, and that the two are not separate. The best way to improve the PER with ASR words is to reduce the WER with better speech to text. This is supported by the fact that the PER - WER (the difference between the two error rates) is relatively constant for the automatic system.

8. DISCUSSION

We have developed an infrastructure for the annotation and scoring of punctuation. Using this infrastructure and scoring technique, we have measured inter-annotator consistency for the RT-S task. We have also developed a preliminary system for the tagging of punctuation, as well as a scoring method using multiple references. This multiple annotator scoring method allows us to differentiate the acceptable variations between human annotations from the errors of the automatic system; whereas they are possibly too close to judge using a single annotator. The combined reference is necessary for punctuation identification, because there are multiple correct punctuations, and because the task of labeling punctuation is hard, even for humans. In the future we need to study the effect of the punctuation and capitalization errors on human readability and fatigue, and preference.

9. REFERENCES

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

[2] Stephanie Strassel, *Simple Metadata Annotation Specification – Version 5.0*, Linguistic Data Consortium, Universitry of Pennsylvannia, 2003.

[3] S. Schwarm J. Kim and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning," in *Proceedings of HLT/NAACL*, 2004, pp. 137–144.

[4] Francis Kubala and Amit Srivastava, *A Framework for Evaluating Rich Transcription Technology*, BBN Ears Website. http://www.speech.bbn.com/ears, 2003.

[5] A. Stolcke D. Hillard M. Ostendorf B. Peskin M. Harper Y. Liu, E. Shriberg, "The icsi-sri-uw metadata extraction system," in *The Proceedings of the International Conference on Spoken Language Processing*, 2004.

[6] Bonnie Door Matthew Snover and Richard Schwartz, "A lexically-driven algorithm for disfluency detection," in *Proceedings of HLT/NAACL*, 2004, pp. 157–160.

[7] Douglas Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas Reynolds, and Marc Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proceedings of Eurospeech*, Geneva, 2003.

[8] Adwait Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *ACL-SIGDAT Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, 1996, pp. 133–142.