Automatic Evaluation of Spoken Summaries for Language Assessment

Anastassia Loukina, Klaus Zechner, & Lei Chen
Educational Testing Service (ETS), Princeton, NJ 08541, USA

Listening. Learning. Leading.®

Motivation

Content accuracy of spoken responses is usually evaluated using a cosine similarity metric between a candidate response and reference response (Attali and Burstein, 2006). This approach requires a large number of reference summaries (Chen, 2013). It also penalizes both the lack of precision and lack of recall.

We investigate whether ROUGE, a popular recall-based metric for the evaluation of automated written summaries, can be applied to the assessment of content accuracy of spoken responses produced by non-native speakers of English.

Data

The speakers were presented with two types of tasks:

1. "... look at a series of six pictures and tell the story that the pictures show ..." (1 question)
2. "... listen to a teacher or a group of students ... talk about what you heard ..." (3 questions)

Scoring rubrics for content accuracy:

Score 4: "... Content is full and appropriate to the task ... although minor errors may occur ..."

Score 3: "... Content is mostly complete and appropriate to the task ... but supporting details and elaboration are limited or lacking ..."

Score 2: "... Development is mostly limited to some (or all) of the main facts, presented one by one. ... Some **key information** may be omitted or inaccurate ... "

Score 1: "... Content is incomplete and/or lacks development ..."

Corpus statistics

- 5,934 spoken responses from 1,611 speakers
- 24 different prompts (4 prompts per speaker)
- Average length of responses: 72 words ($\sigma = 29$)
- ASR WER: 26.5% for picture narration, 29.4% for summarization.

Adapting ROUGE to evaluation of spoken summaries

What is ROUGE?

$$ROUGE_N = \frac{\sum_{Summ \in Reference} \sum_{Ngrams \in Summ} Count_{overlap}(Ngrams)}{\sum_{Summ \in Reference} \sum_{Ngrams \in Summ} Count(Ngrams)}$$

- Developed by Lin and Rey (2004) for the evaluation of automatic text summaries
- Recall-oriented
- Requires only a small number of reference summaries
- Does not need any manual annotation
- Easy and fast to compute automatically
- Successfully used for scoring written test responses (Madnani et al., 2013)

What we did to adapt ROUGE to speech:

- Responses are shorter than automatic summaries (72 words vs. 100 words)
- There are grammatical errors, repetitions, repairs and other disfluencies
- The errors of automatic speech recognition (ASR) introduce further noise

Base ROUGE	New ROUGE	Baseline: CVA
ROUGE-1	ROUGE-1	Cosine similarity (*tf-idf*)
Counts all tokens	Counts types only	Counts types only
$N_{references} = 1$	$N_{references} = 4$	$N_{references} = 4$

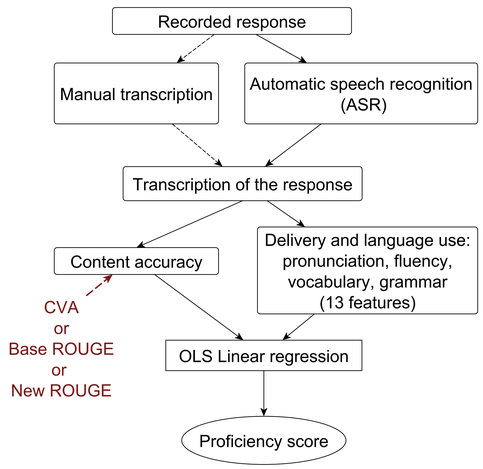
Removing stop-words and lemmatization had no effect on performance.

No further improvement in performance for more than 4 summaries. The choice of reference responses does not matter.

Using ROUGE to score content accuracy

Model	ASR		Manual	
	r	κ	r	κ
Content accuracy only				
CVA only	0.492	0.340	0.469	0.303
Base ROUGE only	0.587	0.440	0.632	0.489
New ROUGE only	0.655	0.540	0.700	0.590
All aspects of proficiency				
Delivery/Lang use only	0.678	0.565	0.678	0.565
D/LU + CVA	0.691	0.600	0.698	0.602
D/LU + Base ROUGE	0.700	0.597	0.719	0.610
D/LU + New ROUGE	0.715	0.617	0.738	0.652

r - Pearson's r between holistic expert human score and predicted score
κ - weighted quadratic kappa between holistic human score and rounded predicted score
Agreement between two expert raters: $k = 0.69$



Is it just the length?

Base ROUGE is very sensitive to the length of the response. The new ROUGE still outperforms CVA if the length of response (N words) is held constant.

	Correlation with human score (r)			
	Absolute		Partial	
	ASR	Manual	ASR	Manual
CVA	0.508	0.451	0.428	0.370
Base ROUGE	0.553	0.589	0.281	0.284
New ROUGE	0.652	0.673	0.478	0.460

Conclusion

- Recall-based ROUGE-1 shows good agreement with expert ratings but is very sensitive to response length.
- The use of types instead of tokens increases the agreement with human ratings and reduces the sensitivity to the response length.
- The use of several reference summaries improves the performance. Only four reference summaries are necessary to achieve reliable results.
- There is only a small drop in performance between human transcriptions and the output of automatic speech recognition.

Selected references

Attali, Y,. and Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3):1–30.

Chen, L. (2013). Applying unsupervised learning to support vector space model based speaking assessment. *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 58–62.

Lin, C.-Y., and Rey, M. (2004). ROUGE: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Madnani, N., Burstein, J., Sabatini, J., and O'Reilly, T. (2013). Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168.

aloukina@ets.org