EVALUATION METRICS FOR SPEECH TRANSLATION

ASHWINI DASARE

Research Scholar, IIIT Dharwad

ashwini@iiitdwd.ac.in

January 8, 2022

Overview

1. Conventional Metrics
   - Standard Evaluation Metrics
2. End-to-End Metrics
   - Limitations and Research Challenges

Standard Evaluation Metrics

- Human Judgement(Subjective)
- Automated Evaluation Metrics(Objective)

The Conventional Approach

- Automatic Speech Recognition[ASR]-Word Error Rate [WER]
- Machine Translation[MT]- Automated Metrics
- Text to Speech Synthesis[TTS]- Mean Opinion Score[MOS]

Word Error Rate

WER =(S+I+D)/N

- S - substitutions
- I- insertions
- D- deletions
- N- number of said words

Word Error Rate

ORIGINAL

Speech to Speech Translation System

$WER = 3/5 = 0.6$

ASR-Output

Speech to *Text*(D) *Conversion*(S) System *Model*(I)

Limitations

- The only Metric for speech to text available
- Works good only on domain specific Language
- Poor on Accents and Homophones

AUTOMATED METRICS

- BiLingual Evaluation Understudy (BLEU)
- Metric for Evaluation of Translation with Explicit word Ordering (METEOR)
- Translation Edit Rate (TER)

- Ref:Condon et al. "Automated metrics for speech translation." In Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems,2009

BLEU

- Developed by IBM researchers in 2001
- BLEU score is to compare n-grams of the candidate with the n-grams of the reference translation
- The more the matches, the better the candidate translation

BLEU-Example

- Kannada sentence:**”Budakattu bhaashegaligaagi Dhwanibhaashaanuvaada ghataka**”
- Reference Sentence: ”**SPEECH TO SPEECH TRANSLATION MODULE FOR TRIBAL LANGUAGES”**
- Candidate Sentence: **”SPEECH TO SPEECH TRANSLATION”**
- 1 gram- $4/4$
- 2 gram- $3/3$
- 3 gram- $2/2$
- 4 gram- $1/1$
- BLEU$=$ Geometric mean of n grams$=$ $\sqrt[4]{(1 * 1 * 1 * 1)}$

BREVITY PENALTY

- BP$=1$ if $c > r$ and BP$=$ $exp(1 - r/c)$ if $c < r$
- BLEU $=$ BP x GM
- exp(1-8/4)$=$ exp(-1)$=$0.37(BLEU score)

BLUE-Limitations

- The matches are position independent
- Measures Precision But Not Recall
- BP is not adding for recall
- Stemming and Synonyms are considered as zero match

The METEOR Measure

- Metric for Evaluation of Translation with Explicit word Ordering
- Developed by Carnegie Mellon University in 2005
- Considers Both Precision and Recall

The METEOR Measure

- Precision (P) $= m/wt$
- Recall (R) $= m/wr$
- Fmean $= 10PR/(R + PR)$
- m: number of unigrams in the candidate found in reference
- wt: Number of unigrams in candidate translation
- wr: Number of unigrams in reference translation

CHUNK PENALTY

- p $= 0.5(C/Um)^3$
- C: Number of chunks in candidate
- Um:Unigrams in candidate

The METEOR-Example

- Reference : SPEECH TO SPEECH TRANSLATION MODULE FOR TRIBAL LANGUAGES
- Candidate : SPEECH TO SPEECH TRANSLATION
- p= $0.5\ (1/4)^3 = 0.0078$
- P= 1, R = 0.5, Fmean = 0.53
- M = Fmean (1- p)=0.53(0.99)= 0.525

TRANSLATION EDIT RATE

- Developed by University Of Maryland in 2006
- TER = (Substitutions + Insertions + Deletions + Shifts)/Reference Words
- Suitable for both Machine and Human Evaluation

MEAN OPINION SCORE

- This is a Human Evaluation technique.
- Measured on the 5-point scale for **Adequacy, Fluency,Naturalness**

Limitations

- Can’t rely on MOS for Speech Translation system
- Synthesized voice does not consider the correctness of translation

Translatotron(2021)

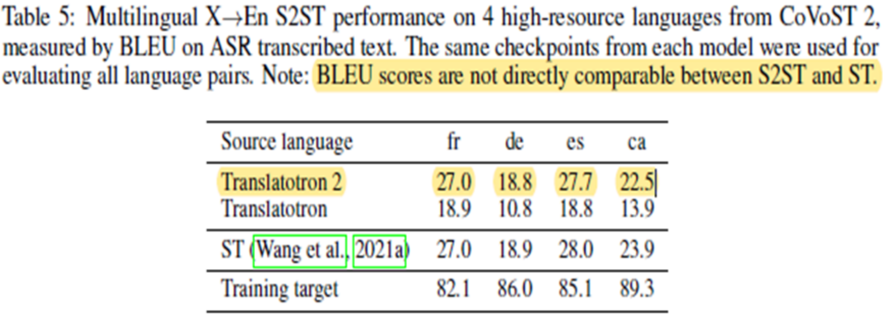

Table 5: Multilingual X→En S2ST performance on 4 high-resource languages from CoVoST 2, measured by BLEU on ASR transcribed text. The same checkpoints from each model were used for evaluating all language pairs. Note: BLEU scores are not directly comparable between S2ST and ST.

Source language	fr	de	es	ca
Translatotron 2	27.0	18.8	27.7	22.5
Translatotron	18.9	10.8	18.8	13.9
ST (Wang et al., 2021a)	27.0	18.9	28.0	23.9
Training target	82.1	86.0	85.1	89.3

Jia, Ye, et al. "Translatotron 2: Robust direct speech-to-speech translation."arXiv preprint arXiv:2107.08661(2021)

Speech Transformer (2021)

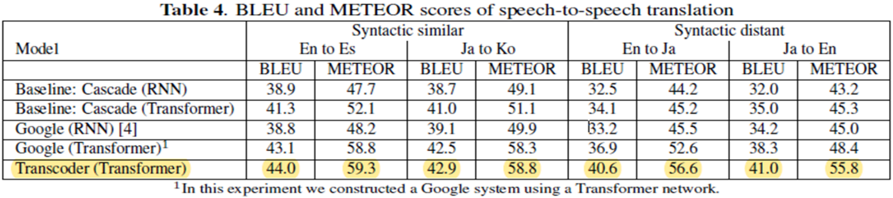

Table 4. BLEU and METEOR scores of speech-to-speech translation

Model	Syntactic similar				Syntactic distant			
	En to Es		Ja to Ko		En to Ja		Ja to En	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Baseline: Cascade (RNN)	38.9	47.7	38.7	49.1	32.5	44.2	32.0	43.2
Baseline: Cascade (Transformer)	41.3	52.1	41.0	51.1	34.1	45.2	35.0	45.3
Google (RNN) [4]	38.8	48.2	39.1	49.9	33.2	45.5	34.2	45.0
Google (Transformer)[1]	43.1	58.8	42.5	58.3	36.9	52.6	38.3	48.4
Transcoder (Transformer)	44.0	59.3	42.9	58.8	40.6	56.6	41.0	55.8

[1] In this experiment we constructed a Google system using a Transformer network.

Kano, Takatomo et al. "Transformer-Based Direct Speech-To-Speech Translation with Transcoder."2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.

Conclusion

- The automated evaluation metrics do not understand Morphological, Semantic and Syntactic factors
- The performance of evaluation metrics depends on human judgments
- No standard Benchmark
- Evaluation Metrics is an Open Challenge

Research Challenges

- Evaluating translation combined with speech synthesis
- Single Metric which can measure the performance of End to End Systems
- Metric for target languages without standardized orthographies
- Effect of dialectal variation/accent variation on Evaluation Metrics

References

- Condon et al. "Automated metrics for speech translation." In Proceedings of the 9th Workshop on Performance Metrics for Intelligent Systems,2009
- Le, Ngoc Tien. "Advanced quality measures for speech translation." PhD diss., Université Grenoble Alpes, 2018.
- Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005
- Jia, Ye, et al. "Translatotron 2: Robust direct speech-to-speech translation."arXiv preprint arXiv:2107.08661(2021).

References

- Kano, Takatomo, Sakriani Sakti, and Satoshi Nakamura. ”Transformer-Based Direct Speech-To-Speech Translation with Transcoder.”2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021.
- Papineni, Kishore, et al. ”Bleu: a method for automatic evaluation of machine translation.”Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002
- Snover, Matthew, et al. ”A study of translation edit rate with targeted human annotation.” Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. 2006.
- https://www.cs.cmu.edu/ alavie/papers/GALE-book-Ch5.pdf
- https://languagelog.ldc.upenn.edu/nll/?p=193

Thank You