Interspeech 2024
1-5 September 2024, Kos, Greece

SeMaScore : A new evaluation metric for automatic speech recognition tasks

Zitha Sasindran, Harsha Yelchuri, T. V. Prabhakar

Department of Electronic Systems Engineering
Indian Institute of Science, Bengaluru, India
{zithas, harshay, tvprabs}@iisc.ac.in

Abstract

In this study, we present SeMaScore, generated using a segment-wise mapping and scoring algorithm that serves as an evaluation metric for automatic speech recognition tasks. SeMaScore leverages both the error rate and a more robust similarity score. We show that our algorithm's score generation improves upon the state-of-the-art BERTScore. Our experimental results show that SeMaScore corresponds well with expert human assessments, signal-to-noise ratio levels, and other natural language metrics. We outperform BERTScore by 41x in metric computation speed. Overall, we demonstrate that SeMaScore serves as a more dependable evaluation metric, particularly in real-world situations involving atypical speech patterns.
Index Terms: Evaluation metric, Speech recognition, Semantic Similarity, Error rate

1. Introduction and motivation

The evaluation of automatic speech recognition (ASR) systems has predominantly centered around the utilization of error-based metrics, such as the word error rate (WER) and character error rate (CER). WER, in particular, has evolved as the de facto metric for benchmarking the performance of ASR models. While state-of-the-art (SOTA) ASR models trained on benchmark speech datasets have consistently demonstrated lower WERs, a notable challenge arises when these well-performing models assess noisy or disordered speech corpora [1], resulting in elevated WER values. Some examples include ASR systems used in real-time applications like medical transcription services applied to patients with speech-impairment or virtual assistants located in highly noisy environments. In these cases the focus needs to shift from quantifying errors to a more thorough assessment of whether the generated hypotheses accurately convey the intended meaning of the spoken sentence.

While WER serves as a valuable metric for quantifying word-level errors in ASR systems, it has limitations [2, 3] in capturing crucial aspects of ASR evaluation , such as semantic similarity and the significance given to each word error in the hypothesis. Consider two scenarios where the WER remains the same: one involving hyphenated words (well-being) represented as a single word (wellbeing) and the other involving misspelt word (velbing). Despite WER being the same in both cases, the former is more closer to the ground truth semantically. This observation implies that generating a hypothesis that perfectly aligns with the ground truth may be of little importance when comprehending the user's intent and the underlying semantic content.

Over the years, numerous research works have sought to provide alternatives to WER [4, 5] as an evaluation metric and have demonstrated its limited alignment with human judgment when applied in the context of spoken conversations [6]. Recent studies [7, 8, 9, 10] have introduced the utilization of pre-trained BERT-based [11] models to compute a semantic distance metric between the ground truth and hypothesis. However, the semantic distance metric calculated at sentence-level overly prioritizes some words over others, making it unreliable. In BERTScore [12], the authors calculated the semantic distance at token-level as an evaluation criterion for ASR systems. However, BERTScore has some limitations as well, and this is presented in Table 1. Table 1 presents two cases where BERTScore fails to provide an accurate score. The first example illustrates two entirely mismatched words, while the second example showcases a sentence with repetitions—a common mistake in ASR due to encoder-decoder model hallucination issues. In both instances, BERTScore assigns a higher similarity score and fails to detect the errors, despite the sentences being flawed. The authors identified this nature of BERTScore and proposed baseline rescaling, which tones down the value of BERTScore. However, we observed that some of the examples still fail after rescaling. For instance, if we consider the same example 1 of Table 1, even after rescaling, the BERTScore drops marginally from 0.98 to 0.86. Further, baseline rescaling involves the computation of BERTScore over a corpus of million sentences for every sub-task in every language. Motivated by these insights, we created a new metric that focuses on both the error rate and a more robust semantic similarity score for the comprehensive evaluation of ASR models.

Table 1: *Examples of cases where BERTScore fails: ground truth (GT) and hypothesis (H) generated.*

GT	H	BERTScore	SeMaScore
Smoking	Something	0.98	0.3
Thank you lord	Thank you thank thank thank lord	0.89	0.62

In this paper, we introduce **SeMaScore**, a evaluation metric for ASR tasks. This innovative metric integrates traditional error rate-based metric with a robust semantic similarity-based approach to evaluate ASR outputs. Furthermore, our findings illustrate that SeMaScore performs effectively even when WER is considerably high when ASR models encounter atypical speech patterns such as disordered, noisy, or accented speech utterances. Our work demonstrates the robustness of the proposed metric in handling transcription errors and its strong alignment with expert human assessments, and other natural language metrics. We show that other metrics fail to correlate well with some scenarios, whereas SeMaScore aligns well with almost all the scenarios. This highlights the versatility and reliability of our metric across a broad spectrum of ASR applications.

10.21437/Interspeech.2024-2033

Table 2: *Example of segment mapping between the ground truth (GT) and hypothesis (H) generated.*

Before mapping	After mapping
GT: I want to have a sandwich	I — want — to — have a — sandwich
H: I vant to havea sand wich	I — vant — to — havea — sand wich

2. SeMaScore

We discuss the methodology for calculating the SeMaScore metric. Our approach consists of four primary phases: Segment mapping, computing segment scores & importance weights, and finally, the calculation of the SeMaScore.

2.1. Segment mapping

Utilizing contextual embeddings of complete sentences from the BERT model to compute sentence similarity, as outlined in [7, 8], naturally prioritizes the keywords within each sentence. BERTScore improves upon this and takes a more complex approach by calculating cosine similarity between every token in both the ground truth (GT) and hypothesis (H), thus finding the mapping between the GT and H. While this token-level granularity is suitable for tasks like summarization and text generation, its application in ASR evaluation is computationally expensive.

ASR evaluation, being a sequential task, does not require the complex token-level mapping used in BERTScore. In our approach, we adopt a sentence alignment technique based on the Levenshtein distance to create a mapping between GT and H. To find the mapping between the sentences, we first align GT and H based on the character-level edit distance. Subsequently, we identify corresponding words or groups of words (segments) between GT and H that align with each other. Table 2 provides an example of this segment mapping approach, where aligned portions of sentences correspond to single words or groups of words. As demonstrated in the table, our approach accommodates cases where words are split (e.g., 'sandwich' $\rightarrow$ 'sand wich'), combined (e.g., 'have a' $\rightarrow$ 'havea'), or misrecognized (e.g., 'have' $\rightarrow$ 'hav'). This adaptable mapping strategy allows us to precisely identify the segment-wise correspondence (GT_M and H_M), between GT and H.

2.2. Computing segment score

After obtaining the mapped segments, GT_M and H_M, the next step is to assess their similarity. In our approach, we extract the token embeddings for each segment in $GT_M = \{GT_M[1], GT_M[2] \cdots GT_M[L]\}$ and $H_M = \{H_M[1], H_M[2] \cdots H_M[L]\}$, from the contextual embeddings of GT and H, i.e., e_{GT} and e_H. Further, we use the mean pooling technique on the token embeddings of a segment to compute a single contextual embedding for each segment i in $GT_M[i]$ and $H_M[i]$ to obtain $e_{GT_M}[i]$ and $e_{H_M}[i]$. Thereby, we ensure that the context of the segments is also considered.

As explained in Section 1, by limiting to cosine similarity between token-level embeddings, as done in BERTScore, has its disadvantages of assigning a higher score. Therefore, we propose the introduction of a penalty for the cosine similarity score by multiplying it with (1 - match error rate (MER)). We leverage MER [6], an error-based metric used to assess character-level mistakes. MER is used instead of the traditional CER, as MER is bounded between 0 and 1.

Algorithm 1: SeMaScore

Data: Ground truth (GT), Hypothesis (H)
Result: $SeMaScore$
$GT_M, H_M \leftarrow$ Get_Mapped_Sentences(GT, H);
$e_{GT}, e_{GT_M}, e_{H_M} \leftarrow$
Get_Segment_Embeddings(GT, H, GT_M, H_M);
for $i \leftarrow 1$ **to** *len*(GT_M) **do**
  $SS_i \leftarrow$ Cos_Sim$(e_{GT_M[i]}, e_{H_M[i]})$;
  $SegScore_i \leftarrow SS_i \times (1 - \text{MER}(GT_M[i], H_M[i]))$;
  $\alpha_i \leftarrow$ Cos_Sim$(e_{GT_M[i]}, e_{GT})$;
  $SeMaScore \leftarrow SeMaScore + \alpha_i \times SegScore_i$;
  $\alpha_{sum} \leftarrow \alpha_{sum} + \alpha_i$;
end
$SeMaScore \leftarrow SeMaScore / \alpha_{sum}$;

2.3. Computing importance weights

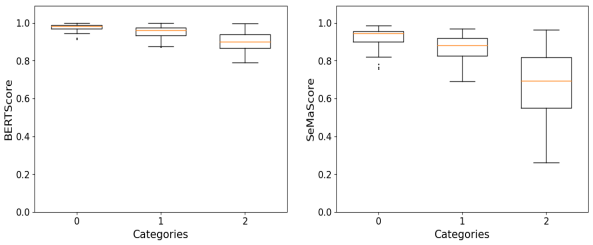
When evaluating a sentence, it is crucial to assign importance to each segment based on its contribution to the meaning of the sentence. For example, as evident in Table 2, the word 'sandwich' carries more significance than the word 'a'. BERTScore addresses this issue by leveraging term frequency-inverse document frequency (tf-idf) to assign weights to words across the ground truth corpus. However, this approach has its limitations. Consider the word 'can,' which can have multiple meanings depending on the context. In the sentence 'I can open the can of soda,' 'can' illustrates the importance of context in understanding its intended sense. Using tf-idf weights alone fails to capture this contextual information, potentially resulting in an inadequate evaluation of the sentence. To effectively determine the importance α_i of segment i, we calculate the cosine similarity between the embedding of the segment ($e_{GT_M}[i]$) and the embedding of the ground truth (e_{GT}). This method allows us to consider the contextual relevance of each segment, providing a more accurate assessment of its contribution to the overall sentence meaning.

2.4. Metric calculation

Algorithm 1 outlines the complete process for calculating SeMaScore. Initially, we apply the segment mapping technique outlined in Section 2.1 to derive the mapped ground truth (GT_M) and hypothesis (H_M). Then, we extract contextual embeddings for each segment and compute the segment score, which is a combination of the similarity score and (1-MER), as elaborated in Section 2.2. Following this, we determine the importance of each segment of GT_M as explained in Section 2.3. Finally, the SeMaScore is computed utilizing all these components, as shown in the Algorithm 1.

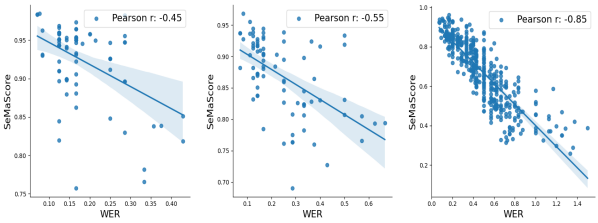
3. ASR Analysis and results

We discuss the diverse settings employed to evaluate our SeMaScore metric. We outline the multiple datasets utilized in our analysis and describe the comparison methodology and results. All results were evaluated against the current SOTA BERTScore. All contextual embeddings were obtained using pre-trained roberta-base model. Revisiting the examples in Table 1 where BERTScore fails, we can see that SeMaScore evaluated the hypotheses better by penalizing them appropriately. All the codes, datasets employed, and results are available from the following link: `https://github.com/zenlab-edgeASR/SeMaScore`.



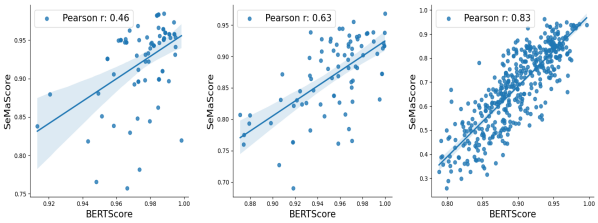
(a) *BERTScore* (b) *SeMaScore*

Figure 1: *Distribution of various metrics for disordered speech*



(a) *Category 0* (b) *Category 1* (c) *Category 2*

Figure 2: *Correlation between SeMaScore and WER for disordered speech*



(a) *Category 0* (b) *Category 1* (c) *Category 2*

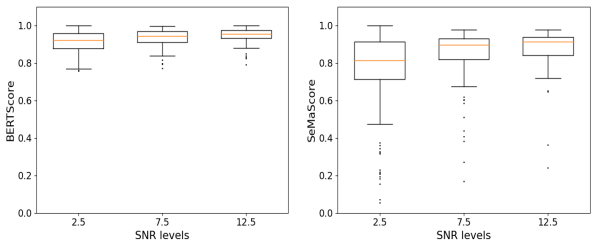
Figure 3: *Correlation between SeMaScore and BERTScore for disordered speech*

3.1. Disordered speech

Our initial set of analyses aims to evaluate the effectiveness of our metric when applied to disordered speech datasets. To achieve this, we employ the Torgo dataset [13], which features speech recordings from individuals with dysarthria. We utilize two distinct models, namely DeepSpeech2 [14] and Wave2vec2 [15], to generate inferences for 600 speech utterances (300 from each model) with an average WER of 0.4. In order to comprehensively assess the metric's performance and utility, we sought the expertise of two language specialists to measure the degree of similarity between the inferred sentences and the ground truth. The assessment criteria we employ, as well as the observed error types, align with those described in [16]. These assessments are categorized into three groups: Category 0: meaning is entirely preserved, Category 1: meaning is mostly preserved, but some errors are present, and Category 2: meaning is entirely different. We utilize this dataset to examine the correlation between the SeMaScore and human assessments.

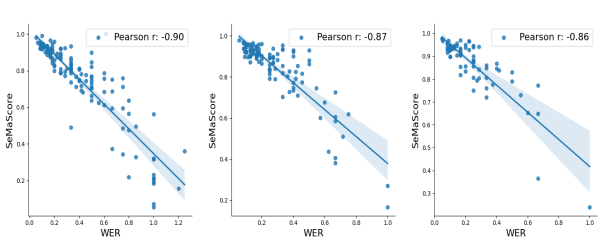
Figure 1 presents the results obtained when BERTScore and SeMaScore are used to evaluate the disordered speech. It is clear from the figures that both BERTScore and SeMaScore correlate with human assessments. From figures 1a and 1b, we can infer that for category 2, BERTScore has a mean of 0.9 with a standard deviation of 0.05 and SeMaScore has a mean of 0.68 with a standard deviation of 0.16 respectively. Since category 2 contains hypotheses whose meaning is different from that of the ground truths, we want the metrics to give a lower score which is well satisfied by SeMaScore.

Figure 2 and Figure 3 depict scatter plots illustrating the correlation between SeMaScore and both WER and BERTScore. The plots show a negative correlation between SeMaScore and WER, as well as a positive correlation between SeMaScore and BERTScore across all categories. From Figure 2, it is apparent that the range of SeMaScore values corresponding to a specific WER value has a wide spread, indicating the consideration of semantic content. Figure 3c demonstrates a strong correlation between our metric and BERTScore. However, SeMaScore produces a sensible range of scores (0.3 to 1) for categories containing semantically diverse sentences, compared to BERTScore (0.85 to 1).



(a) *BERTScore* (b) *SeMaScore*

Figure 4: *Distribution of various metrics for noisy speech*

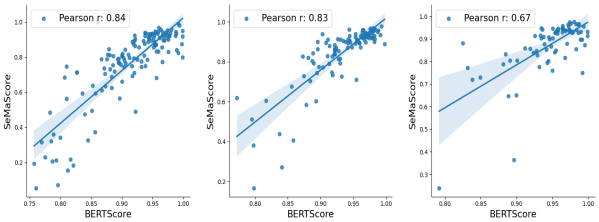


(a) *SNR level 2.5* (b) *SNR level 7.5* (c) *SNR level 12.5*

Figure 5: *Correlation between SeMaScore and WER for noisy speech*

Table 3: *Comparison of BERTScore and SeMaScore with intent accuracy and Named-Entity error rate*

Datasets	WER (%)	BERTScore	SeMaScore	Intent Recognition Accuracy (%)	Named-Entity Error (%)
Set A (ASR Outputs)	27	0.94	0.85	88.16	49.28
Set B	27	0.94	0.87	90.62	21.06
Set C	27	0.90	0.64	81.25	58.91



(a) *SNR level 2.5* (b) *SNR level 7.5* (c) *SNR level 12.5*

Figure 6: *Correlation between SeMaScore and BERTScore for noisy speech*

3.2. Noisy speech

In this experiment, our objective is to investigate the effectiveness of the SeMaScore in evaluating speech utterances under noisy conditions. To achieve this, we make use of the voicebank-demand noisy speech testset [17], which includes speech samples at various signal-to-noise ratio (SNR) levels, including 2.5dB, 7.5dB, and 12.5dB with an average WER of 0.31. We employ the Wave2vec2 model [15] to generate inferences for these noisy speech utterances. Our primary goal is to assess how resilient our metric is to transcription errors in noisy speech and to examine its correlation with the SNR values.

Figure 4 portrays the results discovered when BERTScore and SeMaScore are used to evaluate the noisy speech. The figures show that SeMaScore correlates well with SNR levels. We observe that, at a lower SNR level 2.5, the hypotheses generated correspond to assessment category 2. From figures 4a and 4b, we can infer that for SNR level 2.5, BERTScore has a mean of 0.9 with a standard deviation of 0.06 and SeMaScore has a mean of 0.71 with a standard deviation of 0.22 respectively. Since SNR level 2.5 includes hypotheses whose meaning differs from that of the ground truth, we expect the metrics to produce a lower score, which is again effectively delivered by SeMaScore.

Figure 5 and Figure 6 present scatter plots illustrating the correlation between SeMaScore with WER and BERTScore across various SNR levels. In Figure 5a, at a lower SNR level of 2.5, where the hypotheses may contain more errors, SeMaScore values exhibit a broader distribution similar to WER. In Figure 5b and Figure 5c at higher SNR levels, both WER and SeMaScore values are confined to a lower range, as expected. Overall, WER and SeMaScore correlate well for this task. Similar to our findings in disordered speech experiments, Figure 6 showcases a strong correlation between our metric and BERTScore. SeMaScore produces scores within a broad range (0.1 to 1), in contrast to the higher BERTScore values (0.75 to 1).

3.3. Domain-specific accented speech corpus

In this experiment, we conduct a comprehensive assessment of the SeMaScore in conjunction with other NLU-based metrics, specifically intent recognition accuracy and named-entity error rate, using a domain-specific accented speech dataset. To create this dataset, we curated 500 speech samples from the ATIS text corpus [18]. These speech samples were generated using a text-to-speech system [19] with five distinct accented voices. To evaluate our SeMaScore metric, we initially generated a hypotheses set (referred to as Set A) by feeding these samples through the Deepspeech2 model. To further broaden our evaluation, we created two additional hypotheses sets, namely Set B and Set C, each containing the same WER as Set A. Set B was constructed by introducing errors that have minimal impact on the sentence's overall meaning, such as inserting articles or splitting words in the reference ground truth. Conversely, Set C was generated by introducing errors that included removing or replacing keywords with random words.

Table 3 illustrates the results obtained when we co-evaluate BERTScore and SeMaScore with other NLU-based metrics such as intent recognition accuracy and named-entity error rate. We can observe that BERTScore failed to correlate with both the intent recognition and named entity recognition tasks, whereas SeMaScore correlated with both of them. In case of intent recognition, SeMaScore is directly proportional to intent recognition accuracy, whereas it is indirectly proportional to named-entity error rate in the case of named entity recognition.

3.4. Metric complexity

Consider the scenario where GT consists of m tokens and H contains n tokens. If we consider the time required to calculate cosine similarity as t_{cosine}, then the computation time for the BERTScore metric can be expressed as $n \times m \times t_{cosine}$. Let us assume the time required for edit-distance is given as t_{edit}. The metric calculation time for SeMaScore can be represented as $t_{edit} + 2 \times max(n, m) \times t_{cosine}$ (worst case scenario). We observed that the average metric calculation time for BERTScore on Torgo dataset took 1.95 seconds, whereas SeMaScore took 0.047 seconds. Thus, SeMaScore is 41 times faster than BERTScore.

4. Summary and conclusions

In this work, we introduce a novel evaluation metric for ASR tasks called **SeMaScore**, which utilizes both error rate and a more robust similarity score to evaluate ASR outputs. Our work compares SeMaScore with SOTA BERTScore and demonstrates that it aligns stronger than BERTScore in expert human assessments, SNR levels, and other natural language metrics. Moreover, our approach exhibits lower time complexity, resulting in metric computation times that is 41 times faster than that of BERTScore. These observations highlight the versatility and reliability of our metric across a wide range of ASR applications.

5. References

[1] J. R. Green, R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave, K. Seaver, M. A. Ladewig, J. Tobin, M. P. Brenner *et al.*, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases." in *Interspeech*, 2021, pp. 4778–4782.

[2] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5632–5635.

[3] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz *et al.*, "Automatic human utility evaluation of asr systems: Does wer really predict performance?" in *INTERSPEECH*, 2013, pp. 3463–3467.

[4] I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," IDIAP, Tech. Rep., 2004.

[5] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel *et al.*, "Performance measures for information extraction," in *Proceedings of DARPA broadcast news workshop*. Herndon, VA, 1999, pp. 249–252.

[6] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[7] S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic distance: A new metric for asr performance analysis towards spoken language understanding," *arXiv preprint arXiv:2104.02138*, 2021.

[8] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Evaluating user perception of speech recognition system quality with semantic distance metric," *arXiv preprint arXiv:2110.05376*, 2021.

[9] R. Whetten and C. Kennington, "Evaluating and improving automatic speech recognition using severity," in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, 2023, pp. 79–91.

[10] Z. Sasindran, H. Yelchuri, T. V. Prabhakar, and S. Rao, "Heval: A new hybrid evaluation metric for automatic speech recognition tasks," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[13] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, pp. 1–19, 01 2010.

[14] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[16] J. Tobin, Q. Li, S. Venugopalan, K. Seaver, R. Cave, and K. Tomanek, "Assessing asr model quality on disordered speech using bertscore," *arXiv preprint arXiv:2209.10591*, 2022.

[17] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks." in *Interspeech*, vol. 8, 2016, pp. 352–356.

[18] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in atis?" in *2010 IEEE Spoken Language Technology Workshop*. IEEE, 2010, pp. 19–24.

[19] "Speechelo," *https://speechelo.com/*.