INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation

Janine Rugayan[1], *Giampiero Salvi*[1,2], *Torbjørn Svendsen*[1]

[1]Department of Electronic Systems, NTNU, Norway
[2]KTH Royal Institute of Technology, EECS, Sweden
{janine.rugayan, giampiero.salvi, torbjorn.svendsen}@ntnu.no

Abstract

Automatic speech recognition (ASR) systems have become a vital part of our everyday lives through their many applications. However, as much as we have developed in this regard, our most common evaluation method for ASR systems still remains to be word error rate (WER). WER does not give information on the severity of errors, which strongly impacts practical performance. As such, we examine a semantic-based metric called Aligned Semantic Distance (ASD) against WER and demonstrate its advantage over WER in two facets. First, we conduct a survey asking participants to score reference text and ASR transcription pairs. We perform a correlation analysis and show that ASD is more correlated to the human evaluation scores compared to WER. We also explore the feasibility of predicting human perception using ASD. Second, we demonstrate that ASD is more effective than WER as an indicator of performance on downstream NLP tasks such as named entity recognition and sentiment classification.

Index Terms: ASR evaluation metric, semantic context, user perception

1. Introduction

Automatic speech recognition (ASR) systems have become a vital component of our modern world. For instance, people nowadays use voice-activated virtual assistants to make grocery lists, control household devices, schedule appointments, etc. It is only fitting that we assess the ASR system performance in a way that reflects human understanding and the system's performance on Natural Language Understanding (NLU) or Natural Language Processing (NLP) downstream tasks.

Word error rate (WER), the most widely used evaluation metric, expresses the quality of ASR systems in terms of the word-level edit distance between the reference and ASR hypothesis text normalized by the total number of words in the reference [1]. All the errors are weighed equally, without any regard for the severity of one word error over another. For example, we have someone saying "Find me flights to London". An ASR system transcribes it as "Find *the* flights to London", and another one transcribes it as "Find me flights to *Lisbon*". There is a single word error in both transcriptions but the latter one drastically changes the meaning of the sentence. This illustrates how limited WER can be.

A number of works have proposed modifications to WER in order to combat its limitations. Additionally, others have presented entirely new evaluation metrics as an alternative. Moving from word level accuracy onto the inclusion of semantic information, Semantic-WER (SWER) in [2] used specific weights for insertion, deletion and substitution. Entities and sentiment words are assigned importance weights related to the impact of incorrectly transcribing them. The limitation of SWER is the need for an annotated reference. In addition, domain-specific metrics [3, 4, 5] have been proposed as a means of evaluating ASR systems such that it better reflects the system's performance on various NLU/NLP downstream tasks. However, the disadvantage of these methods is their inability to generalize for various applications.

Meanwhile, transformer-based [6] language models, such as BERT [7] and RoBERTa [8], exhibited the ability to capture semantic information. Thus, representing text as contextualized embeddings was made possible. Reimers *et al.* [9] modified the pretrained BERT and RoBERTa models to generate sentence embeddings and demonstrated their successful application on semantic textual similarity tasks. Likewise, [10, 11] utilized BERT models to evaluate text generation systems (summarization, machine translation, etc.) and presented results which highly correlated with human judgement of text quality. While these works employed the modeled semantic information, they did not focus on the evaluation of ASR system performance.

Kim *et al.* [12] presented an alternative ASR evaluation metric called Semantic Distance (SemDist), which derives sentence-level embedding vectors using RoBERTa and compares them using cosine-similarity. Their results demonstrated that SemDist is a better indicator for various NLU and NLP tasks compared to WER. SemDist was further analyzed in [13] by studying its correlation with the human perception of ASR quality. However, Rugayan *et al.* [14] highlighted that SemDist values are not robust against sentence length due to the averaging of all token embeddings in the sentence. As such, they proposed Aligned Semantic Distance (ASD), which utilizes dynamic programming (DP) to find the optimal alignment between two sequences of token embeddings and calculates semantic closeness as the accumulated distance of the alignment. Their work focused on the evaluation of a Norwegian ASR system and used NorBERT [15] to generate the token embedding sequence for each sentence. They demonstrated that ASD is unaffected by sentence length and illustrated through examples that it provides a more semantically meaningful metric compared to WER.

In this work, we examine the ASD metric against WER in two aspects: correlation with human perception of ASR quality and as an indicator of performance on downstream NLP tasks. Similar to [14], this paper focuses on the evaluation of a Norwegian ASR system. First, we conduct a survey using 30 pairs of reference and ASR transcriptions texts and show that ASD has better correlation to the human perception of ASR transcription quality compared to WER. We also explore the effects of using a monolingual versus a multilingual language model for the ASD metric and perform regression analysis on the results. Furthermore, we build a Gaussian Naive Bayes classifier to pre-

10.21437/Interspeech.2023-1778

dict human evaluation scores and show that using ASD as a feature attains better classification results compared to WER. Finally, we evaluate ASD on Named Entity Recognition (NER) and sentence-level sentiment classification. In contrast to WER, we demonstrate that ASD is a better indicator of downstream NLP task performance and can possibly be used for ASR model selection.

2. Aligned Semantic Distance (ASD)

We use ASD [14] as the semantic-based metric to evaluate the ASR transcriptions. ASD performs token-wise comparison between the embedding vectors for the reference and the hypothesis text. The embedding vectors are derived using a pretrained transformer-based [6] language model. First, the text is tokenized and passed through the transformer model. The embeddings output from all layers of the model is averaged for each token, generating a sequence of token embeddings $e_{\text{ref}}[i]$ and $e_{\text{hyp}}[j]$ for the reference and hypothesis text respectively, as shown in Equation 1.

$$\begin{aligned} E_{\text{ref}} &= \{e_{\text{ref}}[1], e_{\text{ref}}[2], ..., e_{\text{ref}}[N]\} \\ E_{\text{hyp}} &= \{e_{\text{hyp}}[1], e_{\text{hyp}}[2], ..., e_{\text{hyp}}[M]\} \end{aligned} \tag{1}$$

Then, ASD finds the optimal alignment between the reference embedding vector E_{ref} and the hypothesis embedding vector E_{hyp} using dynamic programming (DP). Finally, as shown in Equation 2, the ASD metric is calculated as the minimum accumulated distance of the alignment path D_ϕ, normalized by the reference embedding vector length [14].

$$\text{ASD}(E_{\text{ref}}, E_{\text{hyp}}) \triangleq \min_{\phi} \frac{1}{N} D_\phi(E_{\text{ref}}, E_{\text{hyp}}) \tag{2}$$

It can be observed that ASD depends on the language model used. To explore the language model's effect, we consider two cases. First, we follow [14] and use NorBERT [15], a large-scale monolingual BERT-based language model trained on Norwegian corpora containing both the written standards of Norwegian, Nynorsk and Bokmål. We refer to semantic distances under this setup as *ASD-NorBERT*. For the second case, we utilize the BERT-base multilingual (cased)[1] model [7]. It is pretrained on the top 104 languages with the largest Wikipedia, which includes both Nynorsk and Bokmål. We refer to semantic distances under this setup as *ASD-multiBERT*.

3. Experimental Setup

We examine ASD against WER on two aspects. First, we examine the relationship between the human perception of ASR quality and the evaluation metrics. We hypothesize that a semantic-based metric such as ASD would correlate better to human perception in comparison to WER. Second, we evaluate ASD and WER on downstream NLP tasks. Our assumption is ASD would be more effective than WER as an indicator of task performance.

3.1. Survey on Human Perception of ASR Transcription Quality

We generate a set of ASR transcriptions using a strong baseline end-to-end ASR system [16] with hybrid connectionist temporal classification (CTC) [17] and attention-based encoder-decoder (AED) [18] architecture. The set contains approximately 2600 sentence-like segments, all with manual reference texts from their respective corpora. We use the free speech test set of NB Tale [19], which contains spontaneous narrative monologues, and a part of the test set of Rundkast [20], which includes broadcast radio news shows. Lastly, we take the test set from the Norwegian Parliamentary Speech Corpus (NPSC) [21]. It contains audio recordings of the Q&A sessions of the Norwegian Parliament.

We conduct a survey asking informants to judge the quality of ASR transcriptions selected from the set generated by our baseline ASR system. The questionnaire contains 30 pairs of reference and ASR transcription texts with varying lengths, speech sources, and WERs. The task of the participant is to score the ASR transcriptions from 1 to 5, with 1 being the lowest and 5 being the highest. They are only presented with the reference and ASR transcription texts. ASD's main objective is to measure the semantic distance between the reference and hypothesis text pairs. Therefore, the speech file is not available for playback during the execution of the survey. The guidelines for scoring the ASR transcriptions are as follows:

1. *Bad*: The ASR transcription has several errors that make it completely incomprehensible.
2. *Poor*: The ASR transcription contains errors which make it a lot more difficult to understand. It is not completely unintelligible but means something entirely different compared to the reference.
3. *Fair*: The ASR transcription contains errors which make it a little difficult to understand. It can still be understood to partly mean the same as the reference.
4. *Good*: The ASR transcription contains few errors which makes it a bit different from the reference. It is usually understood to mean the same as the reference.
5. *Excellent*: The ASR transcription contains minor errors or none at all (exact match). It is perfectly comprehensible to mean the same as reference.

We gather a total of 40 participants whose native language is Norwegian. A correlation analysis is performed between the human evaluations scores and the equivalent WERs and ASD values of the survey items. In addition to the language model cases described in section 2, we observe the effects of modifying the layers selected as output of the model. We simply divide the model's 12 layers into 3 groups (bottom, middle, high) and calculate the ASD values using each group. Again, we perform correlation analysis between the ASD values produced by this modification and the human evaluation scores.

3.2. NLP task: Named Entity Recognition, Sentence-level Sentiment Classification

We construct two additional hypotheses sets using the reference text of the baseline ASR transcriptions in subsection 3.1. We have *Worse ASD*, where the ASD mean value is higher than the baseline ASR transcriptions, and *Better ASD*, where the ASD mean value is lower. Both sets have the same mean WER as the baseline ASR transcriptions. To build the Worse ASD set, we take the reference text and randomly *substitute* words with "og" (and). To create minimal disruption in the Better ASD set, we randomly *insert* "og" (and) in the reference text such that all of the key words are retained. The purpose of these additional hypotheses sets is to demonstrate that even if the WER remains constant, ASD values can vary and therefore could potentially be more useful in determining which ASR system is better.

[1] `https://huggingface.co/bert-base-multilingual-cased`

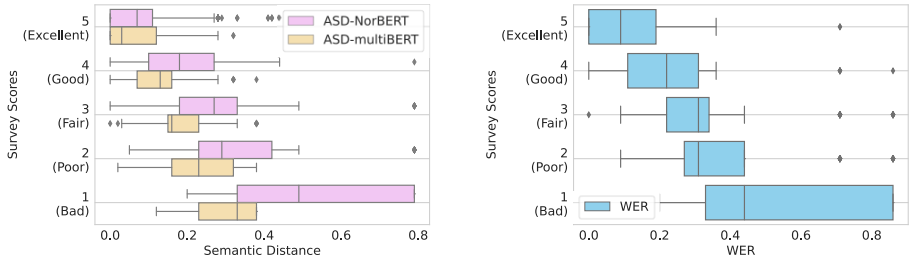

Figure 1: *Distribution of ASD-NorBERT and ASD-multiBERT values (left) and WERs (right) with respect to the human evaluation scores of the ASR transcription quality survey.*

Table 1: *Correlation between survey scores and evaluation metrics. Reported values are Pearson correlation coefficients.*

Metric	Layers	Correlation Coeff.
WER	1 - 12	-0.604
ASD-NorBERT	1 - 12	-0.646
ASD-multiBERT	1 - 12	**-0.683**
ASD-multiBERT	1 - 4	-0.659
ASD-multiBERT	5 - 8	**-0.698**
ASD-multiBERT	9 - 12	-0.684

Table 2: *Comparison of linear regression models using different metric values as input.*

Metric	MSE	MAE	R^2
WER	0.98	0.79	0.34
ASD-NorBERT	0.94	0.78	0.37
ASD-multiBERT	**0.82**	**0.72**	**0.45**

We evaluate the baseline ASR, Worse ASD, and Better ASD hypotheses sets on the downstream NLP tasks of NER and sentence-level sentiment classification. These tasks aid in better understanding of the text. NER classifies each word as one of 10 labels pertaining to persons, organizations, products, events, etc., and sentiment classification tags the sentence polarity either as positive, negative or neutral.

The `nb-bert-base`[2] model [22] is finetuned to perform the NLP tasks. We use the NorNE[3] dataset [23] to finetune the model for NER and the NoReC_sentence[4] dataset [15, 24] to finetune the model for sentence-level sentiment classification. Because our datasets do not have annotations for NER and sentiment classification, we use the finetuned model to annotate the reference text and consider them as our pseudo labels. Then, we perform the NLP tasks on the hypotheses sets and calculate their respective F1-scores.

4. Results and Discussion

4.1. Correlation between Human Perception and ASD

In Table 1, we show the Pearson correlation coefficient between the human evaluation scores and the evaluation metric values for the reference and ASR hypothesis text pairs. It should be noted that more accurate ASR hypotheses result in lower evaluation metric values and in higher human evaluation scores. Due to this inverse relationship, the correlation coefficients are negative.

Results show that semantic-based metrics have better correlation to the human evaluation scores compared to WER. ASD-multiBERT has the strongest correlation at -0.683, outperforming ASD-NorBERT. It is interesting that using a multilingual language model for the ASD metric ended up corresponding better to the scoring task. To understand this phenomenon, we observe the distribution of the evaluation metric values in Figure 1. It shows that ASD-NorBERT values and WERs are more widely distributed compared to ASD-multiBERT. We believe that the more compact distribution of ASD-multiBERT allows it to be more correlated with the discrete scoring levels of the evaluation task. Moreover, it is probable that the difference in distribution of ASD values, especially evident with survey score *1 (Bad)*, is a result of the variation in tokenization between the language models. For example, ASD-NorBERT would tokenize ”forstå” (understand) as a single token but ASD-multiBERT would break it up into ”for” and wordpiece ”##stå”. Wordpieces are subwords learned by the language model during training. ASD-multiBERT tends to divide up a word because it utilizes wordpieces from other languages too. On the contrary, ASD-NorBERT only splits the compound words. We generally observe lower accumulated cosine distances between the aligned token embeddings of wordpieces compared to the token embeddings of whole words.

Since ASD-multiBERT has the strongest correlation to the human evaluation scores, we choose to experiment with it further by modifying the selected output layers of the language model. Table 1 shows that using the middle layers, #5 to 8, achieves the strongest correlation to the human evaluation scores. There have been numerous works which performed probing experiments on BERT [7] in order to understand the way it learns linguistic information. Tenney *et al.* [25] find that earlier parts of the network resolve part-of-speech (POS) tags, and higher layers resolve semantic roles and coreference. Likewise, Jahawar *et al.* [26] show that surface information is encoded in the bottom layers, syntactic and semantic information in the middle layers, and semantic information in the high layers. These findings support the aforementioned experiment results.

4.2. Predicting Human Perception of ASR Quality

We believe that ASD can be utilized for predicting the human perception of ASR quality. We use the 1200 pairs of survey scores and evaluation metric values to perform a regression analysis and to implement a classifier predicting the human evaluation scores of the ASR transcriptions. Table 2 com-

[2] `https://huggingface.co/NbAiLab/nb-bert-base`
[3] `https://github.com/ltgoslo/norne/`
[4] `https://github.com/ltgoslo/norec_sentence`

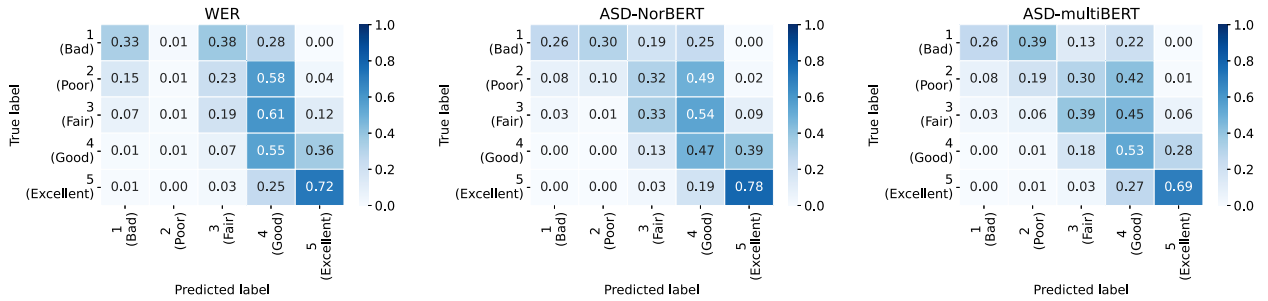

Figure 2: *Normalized confusion matrices of Gaussian Naive Bayes classifier using WER, ASD-NorBERT, and ASD-multiBERT with a 10-fold cross-validation.*

Table 3: *Balanced accuracy scores of Gaussian Naive Bayes classifier using different metric values as the input feature.*

Metric	Accuracy	Standard Dev.
WER	0.361	0.032
ASD-NorBERT	0.390	0.029
ASD-multiBERT	**0.413**	0.035

Table 4: *WER, ASD and F1 scores on the NER & sentiment classification task. Reported ASD values using NorBERT LM.*

			F1 Score	
Hypothesis Set	**WER**	**ASD**	**NER**	**Sentiment**
Baseline	0.074	0.066	0.878	0.938
Better ASD	0.074	0.051	0.944	0.957
Worse ASD (random)	0.074	0.080	0.893	0.898
Worse ASD (entity priority)	0.074	0.093	0.704	0.926

pares the regression models based on different input cases and presents their mean squared error (MSE), mean absolute error (MAE) and coefficient of determination (R^2). Results show that ASD-multiBERT achieves the lowest errors and highest R^2 score.

We implement a Gaussian Naive Bayes classifier using WERs, ASD-NorBERT and ASD-multiBERT values as input feature cases. Since our training data is small, we perform a 10-fold cross-validation. It should be noted that more than half of the transcriptions have a rating of *4 (Good)* and higher. Because of our imbalanced dataset, we calculate the balanced accuracy which is the average of the recall values for each class. Table 3 shows that ASD-multiBERT achieves the highest accuracy. In addition, it attains an MSE of 0.95, which is 24% and 7% lower than the MSE of WER and ASD-NorBERT respectively. From the normalized confusion matrices in Figure 2, we observe that predicted scores are generally higher than the actual ones. This observation could be attributed to the imbalanced dataset. Performance is especially unsatisfactory for *2 (Poor)*, wherein approximately 50% of the time the classifier predicts the examples as *4 (Good)* regardless of the input feature case. Interestingly, WER achieves the best performance for *1 (Bad)* and *4 (Good)*, and ASD-NorBERT for *5 (Excellent)*.

The regression analysis and classification results suggest the feasibility of using ASD to build a prediction model for the human perception of ASR quality. However, independent data is required to verify this.

4.3. NLP downstream tasks and ASD

We evaluate ASD and WER on the NLP tasks of NER and sentence-level sentiment classification. Table 4 shows that while the WER remains constant for all hypotheses sets, the F1-scores for both tasks vary. It illustrates the inability of WER to act as an indicator of NLP task performance. On the other hand, when the ASD value decreases, as in the case of *Better ASD*, the F1-scores for both tasks become higher than the baseline hypothesis set. However, NER F1-score reduction is not generally true when the ASD value increases, as seen in the results for the *Worse ASD (random)* hypothesis set. The reason is most of the words replaced were not named entities. For demonstration purposes, we try prioritizing the replacement of named entities before random words, and report the results under *Worse ASD (entity priority)*. It can now be observed that the NER F1-score is lower than the baseline hypothesis set. These observations indicate that ASD provides better insight on NLP task performance compared to WER and can possibly be used for model selection.

5. Conclusion and Future Work

In this paper, we examine Aligned Semantic Distance (ASD) against WER. First, we gather data on the human perception of ASR transcription quality by conducting a survey with a 5-point evaluation scale. Our correlation analysis show that semantic-based metric ASD has better correlation with human perception compared to WER. We also find that using the BERT-base multilingual (cased) model achieves even better correlation results compared to the monolingual language model NorBERT. In addition, we demonstrate through regression analysis and a Gaussian Naive Bayes classifier that using ASD to build a prediction model for the human evaluation scores is feasible. Finally, we show that ASD is more effective than WER as an indicator of performance on downstream NLP tasks, namely NER and sentence-level sentiment classification.

For our future work, we plan to expand the survey and make it publicly available. In addition, we want to explore incorporating ASD as an optimization criteria in finetuning ASR systems.

6. Acknowledgements

This work was carried out within the EEA and Norway Grants project NORDTRANS - Technology for automatic speech transcription in selected Nordic languages and the Research Council of Norway SCRIBE project (no. 322964).

7. References

[1] 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. [Online]. Available: http://publications.idiap.ch/downloads/reports/2004/rr04-73.pdf

[2] S. Roy, "Semantic-WER: A unified metric for the evaluation of ASR transcript for end usability," *CoRR*, vol. abs/2106.02016, 2021. [Online]. Available: https://arxiv.org/abs/2106.02016

[3] L. van der Werff and W. Heeren, "Evaluating ASR output for information retrieval," *Searching Spontaneous Conversational Speech*, pp. 13–20, 2007.

[4] M. Levit, S. Chang, B. Buntschuh, and N. Kibre, "End-to-end speech recognition accuracy metric for voice-search tasks," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5141–5144, 2012.

[5] M. A. B. Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, "How to evaluate ASR output for named entity recognition?" in *Proc. Interspeech 2015*, 2015, pp. 1289–1293.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410

[10] T. Sun, J. He, X. Qiu, and X. Huang, "BERTScore is unfair: On social bias in language model-based metrics for text generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3726–3739. [Online]. Available: https://aclanthology.org/2022.emnlp-main.245

[11] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 563–578. [Online]. Available: https://aclanthology.org/D19-1053

[12] 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," *Interspeech 2021*, pp. 1977–1981, 2021.

[13] S. Kim, D. Le, W. Zheng, T. Singh, A. Arora, X. Zhai, C. Fuegen, O. Kalinli, and M. Seltzer, "Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric," in *Proc. Interspeech 2022*, 2022, pp. 3978–3982.

[14] J. Rugayan, T. Svendsen, and G. Salvi, "Semantically Meaningful Metrics for Norwegian ASR Systems," in *Proc. Interspeech 2022*, 2022, pp. 2283–2287.

[15] A. Kutuzov, J. Barnes, E. Velldal, L. Øvrelid, and S. Oepen, "Large-scale contextualised language modelling for Norwegian," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 30–40. [Online]. Available: https://aclanthology.org/2021.nodalida-main.4

[16] J. Nouza, P. Červa, and J. Žd'ánský, "Lexicon-based vs. lexicon-free asr for norwegian parliament speech transcription," in *Text, Speech, and Dialogue*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Cham: Springer International Publishing, 2022, pp. 401–409.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[18] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/1068c6e4c8051cfd4e9ea8072e3189e2-Paper.pdf

[19] National Library of Norway. (2015) NB Tale - speech database for Norwegian. [Online]. Available: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-31/

[20] I. Amdal, O. M. Strand, J. Almberg, and T. Svendsen, "RUNDKAST: an annotated Norwegian broadcast news speech corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/486_paper.pdf

[21] National Libary of Norway. (2021) Norwegian parliamentary speech corpus. [Online]. Available: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-58/

[22] P. E. Kummervold, J. De la Rosa, F. Wetjen, and S. A. Brygfjeld, "Operationalizing a national digital library: The case for a Norwegian transformer model," in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021, pp. 20–29. [Online]. Available: https://aclanthology.org/2021.nodalida-main.3

[23] F. Jørgensen, T. Aasmoe, A.-S. Ruud Husevåg, L. Øvrelid, and E. Velldal, "NorNE: Annotating named entities for Norwegian," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4547–4556. [Online]. Available: https://aclanthology.org/2020.lrec-1.559

[24] L. Øvrelid, P. Mæhlum, J. Barnes, and E. Velldal, "A fine-grained sentiment dataset for Norwegian," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5025–5033. [Online]. Available: https://aclanthology.org/2020.lrec-1.618

[25] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452

[26] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. [Online]. Available: https://aclanthology.org/P19-1356