Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh,
Christian Fuegen, Ozlem Kalinli, Michael L. Seltzer

Facebook AI, USA

suyounkim@fb.com

Abstract

Word Error Rate (WER) has been the predominant metric used to evaluate the performance of automatic speech recognition (ASR) systems. However, WER is sometimes not a good indicator for downstream Natural Language Understanding (NLU) tasks, such as intent recognition, slot filling, and semantic parsing in task-oriented dialog systems. This is because WER takes into consideration only literal correctness instead of semantic correctness, the latter of which is typically more important for these downstream tasks. In this study, we propose a novel Semantic Distance (SemDist) measure as an alternative evaluation metric for ASR systems to address this issue. We define SemDist as the distance between a reference and hypothesis pair in a sentence-level embedding space. To represent the reference and hypothesis as a sentence embedding, we exploit RoBERTa, a state-of-the-art pre-trained deep contextualized language model based on the transformer architecture. We demonstrate the effectiveness of our proposed metric on various downstream tasks, including intent recognition, semantic parsing, and named entity recognition.

Index Terms: ASR evaluation metric, spoken language understanding, natural language understanding.

1. Introduction

While the adoption of Word Error Rate (WER) as the de facto evaluation metric has served to advance automatic speech recognition (ASR) research over the decades, there has been an increasing interest in the speech recognition community to consider a more suitable evaluation measure for downstream Natural Language Understanding (NLU) applications, such as intent recognition, slot filling and semantic parsing for task-oriented dialog. This is primarily because WER has been shown to have limitations in measuring semantic correctness, as it is derived from the word-level edit distance between the true transcription and the ASR hypothesis, where every error (substitution, insertion, or deletion) is weighted equally. For example, if the reference is "*This is a cat*" and two ASR systems generate different hypotheses: "*This is the cat*" and "*This is a cap*", then the former system would be preferred by a downstream NLU system. However, WER by itself cannot identify which system is better as the error rates are identical (one substitution error). Past research has highlighted such limitations of WER and demonstrated that improvements in NLU can be obtained while observing a worse WER [1–3].

Motivated by the limitations of WER, alternative measures have been proposed. [4] presented word information preserved (WIP) based on mutual information between the reference and the hypotheses. [5] proposed a new measure that includes named Entity Error Rate (EER), and the stop-word-filtered WER, for taking word importance weight into account.

[6–8] attempted to adopt information retrieval to measure the performance. While these metrics addressed some of WER's limitations, all of them are still based on the literal-level word correctness and do not allow for direct analysis of performance at the semantic level of the sentence. A recent study [9] have attempted to optimize ASR using NLU component jointly, however, it requires the dataset consisting of audio recordings and corresponding NLU annotations, where it may be unavailable or limited.

Recently, substantial work has shown that pre-trained neural language models, trained on billions of words, can learn universal language representations of text in the form of low-dimensional continuous feature vector (i.e., embedding) in the semantic space. These embeddings can then be plugged into a variety of downstream tasks, such as textual similarity, question answering, paraphrasing, sentiment analysis, etc., to drastically improve their performance [10–12]. In 2017, [10] introduced ELMo and demonstrated that contextualized word representations from this model outperformed earlier word embeddings such as Word2Vec [13] and GloVe [14] by capturing linguistic context in addition to word-level syntax and semantics. The GPT model [15] proposed a new architecture using transformers [16] and was used for text generation tasks. BERT, which is based on bidirectional transformer, set a new state-of-the-art performance on 11 NLU tasks [11]. Later, RoBERTa subsequently showed that BERT can be further improved by robustly optimized pre-training process [12]. More importantly, BERT and RoBERTa have demonstrated their ability to derive semantically meaningful sentence embeddings that can be compared using cosine similarity [17]. To the best of our knowledge, there have been no studies thus far that leverage these models to evaluate the performance of ASR systems.

In this work, we propose a novel Semantic Distance (SemDist) measure as an alternative performance metric for ASR systems to capture semantic correctness. We define Semantic Distance as the distance between the reference and an ASR hypothesis in the sentence embedding semantic space. To represent the reference and hypothesis as a sentence embedding, we exploit RoBERTa [12], a state-of-the-art pre-trained deep contextualized language model based on the transformer architecture. We evaluate SemDist on several downstream tasks, including intent recognition, semantic parsing, and named entity recognition. We demonstrate that our proposed metric has better correlation with NLU performance than WER and can potentially be used as part of the model selection process.

2. Semantic Distance

In this section, we describe our proposed SemDist as an alternative ASR performance metric. Figure 1 illustrates the overall procedure to obtain the SemDist from the ASR systems A and
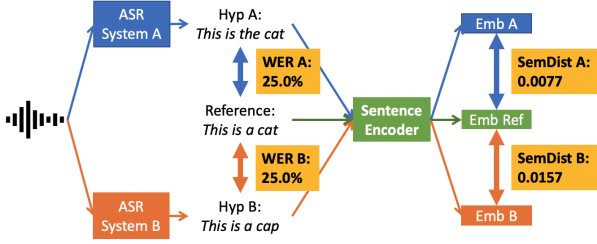
Figure 1: *Two metrics for evaluating ASR systems: the WER and our proposed SemDist.*

Table 1: *Example comparison of our proposed SemDist and other conventional metrics, WER, NER, and POS tagging accuracy of two hypotheses from different ASR systems. The reference transcription is "This is a cat."*

ASR	Hypo.	WER	NER	POS	SemDist
A	*This is the cat*	25.0%	None	100%	0.0077
B	*This is a cap*	25.0%	None	100%	0.0157

B in addition to the WERs. Our proposed SemDist is calculated in two steps. First, we exploit pre-trained sentence embedding models to map the utterances into a sentence embedding space (in Section 2.1). Second, we measure semantic distance by using the cosine similarity function (in Section 2.2).

2.1. Sentence Embeddings

To represent the reference and hypothesis in the sentence-level semantic embedding space, we use RoBERTa [12], a state-of-the-art pre-trained masked language model. It uses the same architecture as BERT [11], with the robustly optimized training method, (i.e. longer training, with bigger batches, dynamically changing the masking pattern, etc), described in [12]. It has produced state-of-the-art results on a wide variety of challenging NLP benchmarks, such as GLUE, SQUAD and RACE. RoBERTa/BERT employ bidirectional training of transformers [16], allowing the models to learn a deeper sense of language context. Further, with the language masking strategy used for training, the model learns to predict intentionally masked sections of text. Most importantly, open source RoBERTa and BERT models, pre-trained on billions of words, are readily available and allow fast fine-tuning with small modifications (i.e. additional output layer) on any specific final task. This form of transfer learning, where pre-trained models are used as starting points for task-specific models, has shown significant breakthroughs in semantic textual similarity tasks [18].

In our method, we pass the reference and hypothesis through the pre-trained RoBERTa and perform a pooling operation by computing the mean of all output vectors. The pre-trained model architecture that we used in this work is BERT$_{\text{BASE}}$ [11], which has 12 transformer layers with 768 hidden size and 12 self-attention heads, for a total of 110M parameters. Thus, a single 768-dimensional sentence embedding vector is computed for each reference and hypothesis.

2.2. Cosine Distance Scoring

In the sentence embedding space, a simple cosine distance has been applied successfully to compare two utterances for semantic textual similarity decision [17]. Given two sentence embeddings: an embedding of the reference transcription, e_{ref}, and an embedding of the hypothesis generated from an ASR system, e_{hyp}, SemDist is calculated as follows:

$$\texttt{SemDist}(e_{\text{ref}}, e_{\text{hyp}}) = 1 - \frac{(e_{\text{ref}})^T \cdot e_{\text{hyp}}}{||e_{\text{ref}}|| \cdot ||e_{\text{hyp}}||} \qquad (1)$$

Note that the cosine distance only considers the angle between the two sentence embeddings and not their magnitudes. SemDist is bounded between 0 and 1, where lower scores indicate higher semantic similarity and vice versa.

Table 1 demonstrates the difference between our proposed SemDist and other conventional metrics, WER, Named Entity Recognition F1-score (NER), POS tagging accuracy (POS) on two different ASR hypotheses, given the reference transcription "*This is a cat*". Naturally, the downstream tasks prefer Model A to Model B because Model A is more semantically correct. As seen in this example, WER and other metrics cannot separate these two models since it only measures literal word-level correctness. However, SemDist can indicate that Model A (0.0077) performed better than Model B (0.0157).

3. Experimental Setup

3.1. Overall Experiment Pipeline

In order to demonstrate the effectiveness of our proposed SemDist metric on various realistic large-scale downstream tasks, we use strong baseline ASR and NLU systems, and evaluate on our large-scale in-house ASR/NLU dataset. We describe these baseline systems and evaluation dataset in more detail in Section 3.2, Section 3.4, and Section 3.3, respectively.

In order to address our main research question "*Can SemDist identify which ASR system is better even when WER is same?*", we derive three more ASR outputs, hypotheses sets, in addition to our ASR baseline output. We then evaluate and compare the performance of these hypotheses set in NLU task by using NLU metrics (in Section 3.6).

We first obtain the hypotheses (Set A) from our strong ASR baseline on the evaluation dataset. We then generate Set B that has the exactly same WER, but has worse (higher) SemDist. To do so, based on each hypothesis' number of substitution/insertion/deletion errors in Set A, we substitute or insert the true/reference word with a random word or delete the random position of the reference word. On the contrary, we generate Set C that has the exactly same WER, but has better (lower) SemDist. To do so, we change the order of two random reference word or add the articles (i.e. '*a*' or '*an*') to minimize the damage of the meaning of the reference sentence. Finally, we also generate Set D that has better (lower) SemDist without limiting to have same WER, but also without artificial way. To do so, we use an external neural language model trained on unpaired text data and perform the shallow fusion [19] with internal LM subtraction [20] on top of our strong ASR baseline. In this way, we can obtain hypotheses with better SemDist, better WER and more realistic insight on SemDist.

3.2. ASR Task

We first build a strong baseline ASR system by employing a large-scale in-house ASR training dataset consisting of two parts. The first part comprises of 1.7M hours of English video data publicly shared by Facebook users; all videos are completely de-identified, and both transcribers and researchers do not have access to any user-identifiable information (UII). The

second part contains approximately 50K hours of manually transcribed de-identified data with no UII in the voice assistant domains.

Our ASR model is an end-to-end sequence transducer, a.k.a. RNN-T [21] with approximately 83M total parameters. The acoustic encoder is a 20-layer streamable low-latency Emformer model [22] with a stride of 6, 60ms lookahead, 300ms segment size, 512-dim input, 2048-dim hidden size, eight self-attention heads, and 1024-dim fully-connected (FC) projection. The predictor consists of three Long Short Term Memory (LSTM) layers with 512-dim hidden size, followed by 1024-dim FC projection. The joiner network contains one Rectified Linear Unit (ReLU) and one FC layer. The target units are 4095 unigram WordPieces [23] trained with SentencePiece [24]. The model is first trained for 4 epochs using sub-word regularization ($l = 5$, $\alpha = 0.25$) [23], SpecAugment LD policy [25], and AR-RNNT loss [26] (left buffer 0, right buffer 15), where the alignment is provided by a chenone hybrid acoustic model (AM) [27]. Finally, we fine-tune the model for 1 epoch with trie-based deep biasing [28].

3.3. Evaluation Dataset

Our in-house annotated evaluation sets for ASR task have two main domains: open-domain dictation and assistant-domain voice commands. The open-domain dictation set includes 22.9K de-identified utterances (305K words) collected from crowd-sourced workers on mobile devices. It contains a mix of short-form (9.3 words/utterance) and long-form (19.0 words/utterance) dictation data under diverse recording environments. The assistant-domain voice commands set includes 15K manually transcribed de-identified utterances (46K words) collected from voice activity of volunteer participants. The participants consist of households that have consented to have their voice activity reviewed and analyzed. We use these datasets to analyze the relationship between WER and SemDist, and basic NLP metrics (in Section 4.1 and Section 4.2).

For NLU, annotated evaluation sets have only assistant-domains, and there are 10k utterances that overlap with the ASR assistant-domain evaluation set. We use this dataset (NLU ∩ ASR) for the tasks of intent recognition and semantic parsing (in Section 4.3).

Table 2: *Description of ASR/NLU task evaluation dataset*

Task	Domain	# utter	# word	Avg. Len.
ASR	Open	23k	305k	13.3
ASR	Assistant	15k	46k	3.0
NLU ∩ ASR	Assistant	10k	25k	2.4

3.4. NLU Task: Intent Recognition, Semantic Parsing

For NLU downstream task, we evaluate four sets of assistant-domain hypotheses (A/B/C/D) with different SemDist on Intent Recognition and Semantic Parsing [29] tasks. Intent recognition is a text classification task where we predict the top-level intent of the utterance from a set of 351 intent types. For the semantic parsing task, we use the recently introduced decoupled semantic representation form [30]. The decoupled representation allows for compositional semantic structures, where a slot can further contain nested intents and slots within itself, providing high expressiveness for task-oriented dialog systems. Figure 2 shows an example of the decoupled representation for the the utterance "*Please remind me to call John*", which has
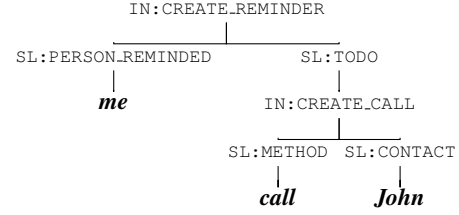
IN:CREATE_REMINDER as the the top-level intent.

Figure 2: *Decoupled semantic representations for the single utterance "Please remind me to call John".*

We build a strong baseline NLU system by using an internal dataset, which consists of about 475k annotated utterances, across 38 domains and 351 intents. The utterances in this dataset were generated via crowd-sourcing and were manually labelled by annotators using the process described in [29].

Our NLU model is a sequence-to-sequence architecture, described in [30], where the source sequence is the utterance and the target sequence is the serialized decoupled representation. At every decoding step, the model can either generate a token from the intent-slot ontology, or copy a token from the source sequence via a pointer-generator mechanism. The model uses two distinct stacked bidirectional LSTMs [31] as the encoder and stacked unidirectional LSTMs as the decoder. Both consist of two layers of size 512, with randomly initialized embeddings of size 300. We also incorporate contextualized word vectors, by augmenting the input with ELMo embeddings [10].

3.5. NLP Task: Named Entity Recognition

In addition to the above NLU tasks, we also evaluate four sets of open-domain hypotheses (A/B/C/D) with different SemDist on a Named Entity Recognition (NER) task. Recognition of named entities such as names of people, organizations, locations, etc, is often used to understand the meaning of text. Thus, we investigate how SemDist relates to the performance on a NER task. Since our dataset does not have annotated entities, we use an open-source software library Spacy [32] to generate their predefined 18 different entities from the reference transcriptions and use them as pseudo labels. We then generate the entities for each hypotheses set and measure the F1-score of the entities.

3.6. Metrics for Downstream Tasks

To evaluate the NLP/NLU performance in the downstream task, we used four metrics:

1. **Intent accuracy (IntentAcc)**: Percentage of utterances where the top-level intent in the decoupled form in the prediction matches the ground truth.

2. **Exact match accuracy (EM)**: Similar to [29], we define exact match accuracy as the percentage of utterances where the complete decoupled form is correct. This is the strictest metric, which is 1 only when all the intents and the slots in the utterance are predicted correctly.

3. **Exact match tree accuracy (EM Tree)**: One drawback of the EM metric is that it will always be 0 when ASR makes a mistake in recognizing slot tokens. Therefore, to study the effectiveness of NLU in the light of such mistakes, we also evaluate the exact match accuracy of the decoupled form after dropping the slot text, which allow us to identify the percentage of utterances where

NLU was able to identify the correct semantic frame, regardless of ASR errors in recognizing slot tokens.

4. **NER-F1**: F1-score of the predicted named entities of the ASR hypotheses

4. Results and Discussions

4.1. Correlation between WER and Semantic Distance

We first analyze the correlation between our proposed SemDist and WER. As seen in Figure 3, we observe that SemDist and WER are highly positively correlated in both open and assistant domains. The Pearson correlation coefficients are 0.72 and 0.65 in the open and assistant domain respectively. In addition, we observe that as WER gets higher it shows more widely spread SemDist at the same WER. This suggests that "not all errors are equal," and by focusing exclusively on WER we may miss more nuanced differences between the hypotheses.
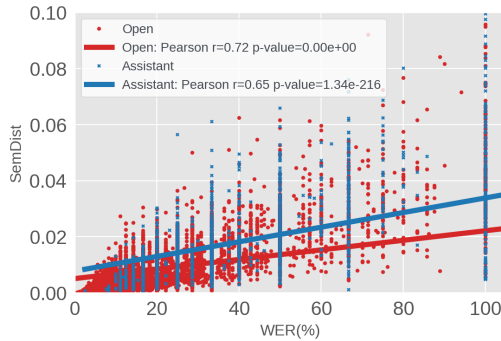

Figure 3: *Correlation between SemDist and WER. The red 'o' marks represent 7,871 utterances out of the entire 23k open-domain testset that show 0 < WER <= 100%. The blue 'x' marks represent 1,829 utterances out of the entire 10k assistant-domain testset that show 0 < WER <= 100%*

4.2. Open-domain: NER and Semantic Distance

Table 3: *Results of WER, SemDist, and NER F1 score*

	WER	SemDist	NER-F1
Set A (BS)	7.44	0.0033	0.747
Set B (WorseSem)	7.44	0.0044	0.590
Set C (BetterSem)	7.44	0.0028	0.846
Set D (BS+LM)	7.03	0.0031	0.758

We next investigate the relationship between the SemDist and the Named Entity Recognition (NER) on the open-domain test set (described in 3.5), which has 23k utterances. Table 3 shows the WER, SemDist, and NER F1-score of four different sets of ASR hypotheses: A(BS)/B(WorseSem)/C(BetterSem)/D(BS+LM) (described in 3.1) on the evaluation set. We observed that as SemDist reduces entity F1-score increases, even with the same WER (Set A vs. Set C). The result indicates that our proposed SemDist measure also aligns with the simple NLP task of NER. Note that the Entity Error Rate(EER) is often used as an additional metric of ASR performance; however, it still has limitations in measuring semantic correctness as seen in Table 1.

4.3. Assistant-domain: NLU tasks and Semantic Distance

We also analyzed the relationship between SemDist and the NLU task (in 3.4) on the assistant-domain test set, which has 10k utterances. Similar to the NER experiments (in 4.2), we compared NLU metrics of the four different sets of ASR hypotheses: A(BS)/B(WorseSem)/C(BetterSem)/D(BS+LM) (described in 3.1) by using our NLU system, described in 3.4. Table 4 shows the WER, SemDist, and NLU metrics: Intent accuracy, EM, and EM Tree (described in 3.6). We observed that as SemDist reduces, Intent accuracy, EM, and EM Tree increase, even with the same WER (Set A vs. Set C). The results indicate that our proposed SemDist can be a better indicator than WER for various downstream NLU tasks as well.

Table 4: *Results of WER, SemDist, and NLU Metrics*

	WER	SemDist	IntentAcc	EM	EM Tree
Set A (BS)	6.16	0.0024	94.63	90.81	91.34
Set B (WorseSem)	6.16	0.0030	94.28	90.27	90.73
Set C (BetterSem)	6.16	0.0017	96.22	92.98	93.08
Set D (BS+LM)	6.01	0.0023	94.84	91.14	91.58

4.4. Examples

Table 5 shows example with their SemDist on the open-domain set. We selected the examples that have same WER on both Set A and D. In the first example, although both hypotheses A and D are incorrect and has same WER, SemDist indicates that D is more semantically close to REF. Since (*'she' vs. 'he'*) are at least the same Part-Of-Speech(POS) tag - subject, we expect that our SemDist may be beneficial to the downstream tasks, such as sentence parsing, POS tagging. In the second example, even though A is more similar in pronunciation (*'hitting that' vs. 'had not'*), we observed that SemDist is higher in A because it is contradict the REF. We also observed that our SemDist takes semantically more important word (*'aw/or' vs. 'aw/oh'*) into account on measuring as seen in the third example.

Table 5: *Examples of hypothesis with SemDist.*

SemDist	Examples
	REF: *she is so cute*
0.0112	**A:** *heat is so cute*
0.0031	**D:** *he is so cute*
	REF: *we **hitting that** new **club** tonight girl*
0.0219	**A:** *we **had not** new **clubs** tonight girl*
0.0167	**D:** *we **had** new **clubs** tonight girl*
	REF: *aw you are all so sweet*
0.0057	**A:** *or you are all so sweet*
0.0050	**D:** *oh you are all so sweet*

5. Conclusion and Future Work

In this work, we propose a novel Semantic Distance (SemDist) as an alternative evaluation metric for ASR systems, capable of measuring the semantic correctness. The SemDist measures the semantic distance between the reference and hypothesis in the embedding space by using the state-of-the-art pre-trained deep contextualized language model, RoBERTa. We demonstrate the effectiveness of our metric on various NLP downstream tasks, including named entity recognition, intent recognition, and semantic parsing. In future, we plan to explore how our Semantic Distance can be used to train ASR systems, as an additional objective.

6. References

[1] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proc. ASRU*. IEEE, 2003, pp. 577–582.

[2] J. S. Garofolo, C. G. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story." *NIST SPECIAL PUBLICATION SP*, vol. 500, no. 246, pp. 107–130, 2000.

[3] D. Grangier, A. Vinciarelli, and H. Bourlard, "Information retrieval on noisy text," IDIAP, Tech. Rep., 2003.

[4] 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.

[5] J. S. Garofolo, E. M. Voorhees, C. G. Auzanne, V. M. Stanford, and B. A. Lund, "1998 TREC-7 spoken document retrieval track overview and results," *NIST SPECIAL PUBLICATION SP*, pp. 79–90, 1999.

[6] 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.

[7] M. J. Hunt, "Figures of merit for assessing connected-word recognisers," *Speech Communication*, vol. 9, no. 4, pp. 329–336, 1990.

[8] 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.

[9] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," in *ICASSP*. IEEE, 2018, pp. 5754–5758.

[10] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *NAACL*, 2018.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[12] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[15] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[18] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 5754–5764. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[19] S. Kim, Y. Shangguan, J. Mahadeokar, A. Bruguier, C. Fuegen, M. L. Seltzer, and D. Le, "Improved Neural Language Model Fusion for Streaming Recurrent Neural Network Transducer," in *Proc. ICASSP*, 2021.

[20] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, and Y. Gong, "Internal Language Model Estimation for Domain-Adaptive End-to-End Speech Recognition," in *Proc. SLT*, 2021.

[21] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML Representation Learning Workshop*, 2012.

[22] Y. Shi, Y. Wang, C. Wu, C. Yeh, J. Chan, F. Zhang, D. Le, and M. L. Seltzer, "Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition," in *Proc. ICASSP*, 2021.

[23] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates," in *ACL*, 2018.

[24] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proc. EMNLP: System Demonstrations*, 2018.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[26] J. Mahadeokar, Y. Shangguan, D. Le, G. Keren, H. Su, T. Le, C. Yeh, C. Fuegen, and M. L. Seltzer, "Alignment Restricted Streaming Recurrent Neural Network Transducer," in *Proc. SLT*, 2021.

[27] D. Le, X. Zhang, W. Zheng, C. Fuegen, G. Zweig, and M. L. Seltzer, "From Senones to Chenones: Tied Context-Dependent Graphemes for Hybrid Speech Recognition," in *Proc. ASRU*, 2019.

[28] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, "Deep Shallow Fusion for RNN-T Personalization," in *Proc. SLT*, 2021.

[29] S. Gupta, R. Shah, M. Mohit, A. Kumar, and M. Lewis, "Semantic parsing for task oriented dialog using hierarchical representations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 2787–2792. [Online]. Available: https://doi.org/10.18653/v1/d18-1300

[30] A. Aghajanyan, J. Maillard, A. Shrivastava, K. Diedrick, M. Haeger, H. Li, Y. Mehdad, V. Stoyanov, A. Kumar, M. Lewis, and S. Gupta, "Conversational semantic parsing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 5026–5035. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.408

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[32] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303