1. SEABORN K, MIYAKE N P, PENNEFATHER P, et al. Voice in human–agent interaction: a survey[J]. ACM Computing Surveys, 2021, 54(4): 1-43.
2. 中泰证券研究所. Cerence：人工智能助力未来出行—车智能化海外公司系列报告[EB/OL]. (2021-04-25) [2022-12-25]. https: pdf.dfcfw.com/pdf/H3\_AP202104261487921362\_1.pdf?1619458440000.pdf.
3. 陈艳华. 基于智能交互的车载语音系统的设计与实现[D]. 北京: 北京交通大学, 2020: 1-2.
4. DERIU J, RODRIGO A, OTEGI A, et al. Survey on evaluation methods for dialogue systems[J]. Artificial Intelligence Review, 2021, 54(1): 755-810.
5. RIGOLL G, NEUKIRCHEN C. A new approach to hybrid HMM/ANN speech recognition using mutual information neural networks[C]. 10th Conference on Neural Information Processing Systems, Denver, USA, 1996: 772-778.
6. MOHAMED A, DAHL G, HINTON G. Deep belief networks for phone recognition[C]. 23rd Conference Neural Information Processing Systems Workshop on Deep Learning for Speech Recognition and Related Applications, Vancouver, Canada, 2009: 39-39.
7. ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Convolutional neural networks for speech recognition[J]. IEEE-ACM Transactions on Audio, Speech, And Language Processing, 2014, 22(10): 1533-1545.
8. GRAVES A. Connectionist temporal classification[M]. Springer, Berlin, Heidelberg: Supervised Sequence Labelling with Recurrent Neural Networks, 2012: 61-93.
9. SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]. 28th Conference Neural Information Processing Systems, Montreal, Canada, 2014: 3104-3112.
10. GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, USA, 2006: 369-376.
11. GRAVES A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012.
12. GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]. 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 2013: 6645-6649.
13. GRAVES A. Generating sequences with recurrent neural networks[J]. arXiv preprint arXiv:1308.0850, 2013.
14. BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
15. CHOROWSKI J, BAHDANAU D, CHO K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: first results[J]. arXiv preprint arXiv:1412.1602, 2014.
16. BAHAR P, ZEYER A, SCHLÜTER R, et al. On using 2d sequence-to-sequence models for speech recognition[C]. 44th IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 2019: 5671-5675.
17. VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. 31st Conference Neural Information Processing Systems, Long Beach, USA, 2017: 5998-6008.
18. ZHOU S, DONG L, XU S, et al. A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin Chinese[C]. 25th International Conference on Neural Information Processing, Siem Reap, Cambodia, 2018: 210-220.
19. MERBOLDT A, ZEYER A, SCHLÜTER R, et al. An analysis of local monotonic attention variants[C]. 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 2019: 1398-1402.
20. ZEYER A, IRIE K, SCHLÜTER R, et al. Improved training of end-to-end attention models for speech recognition[J]. arXiv preprint arXiv:1805.03294, 2018.
21. GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.
22. QIN L, XIE T, CHE W, et al. A survey on spoken language understanding: recent advances and new frontiers[J]. arXiv preprint arXiv:2103.03095, 2021.
23. RAVURI S, STOLCKE A. Recurrent neural network and LSTM models for lexical utterance classification[C]. 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, 2015: 35-40
24. YAO K, ZWEIG G, HWANG M Y, et al. Recurrent neural networks for language understanding[C]. 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013: 2524-2528.
25. MESNIL G, HE X, DENG L, et al. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding[C]. 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013: 3771-3775.
26. MESNIL G, DAUPHIN Y, YAO K, et al. Using recurrent neural networks for slot filling in spoken language understanding[J]. IEEE-ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(3): 530-539.
27. ZHANG X, WANG H. A joint model of intent determination and slot filling for spoken language understanding[C]. 25th International Joint Conference on Artificial Intelligence, New York, USA, 2016: 2993-2999.
28. HAKKANI-TÜR D, TÜR G, CELIKYILMAZ A, et al. Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM[C]. 17th Annual Conference of the International Speech Communication Association, San Francisco, USA, 2016: 715-719.
29. GOO C W, GAO G, HSU Y K, et al. Slot-gated modeling for joint slot filling and intent prediction[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, USA, 2018: 753-757.
30. LI C, LI L, QI J. A self-attentive model with gate mechanism for spoken language understanding[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 3824-3833.
31. Qin L, Che W, Li Y, et al. A stack-propagation framework with token-level intent detection for spoken language understanding[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019: 2078-2087.
32. NIU P, CHEN Z, SONG M. A novel bi-directional interrelated model for joint intent detection and slot filling[J]. arXiv preprint arXiv:1907.00390, 2019.
33. ZHANG C, LI Y, DU N, et al. Joint slot filling and intent detection via capsule neural networks[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 5259-5267.
34. LIU Y, MENG F, ZHANG J, et al. CM-Net: a novel collaborative memory network for spoken language understanding[C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019: 1051-1060.
35. ZHANG L, MA D, ZHANG X, et al. Graph LSTM with context-gated mechanism for spoken language understanding[C]. Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 9539-9546.
36. Qin L, Liu T, Che W, et al. A co-interactive transformer for joint slot filling and intent detection[C]. 46th IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021: 8193-8197.
37. DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, USA, 2019: 4171-4186.
38. CHEN Q, ZHUO Z, WANG W. Bert for joint intent classification and slot filling[J]. arXiv preprint arXiv:1902.10909, 2019.
39. CHEN H, LIU X, YIN D, et al. A survey on dialogue systems: recent advances and new frontiers[J]. ACM SIGKDD Explorations Newsletter, 2017, 19(2): 25-35.
40. MCTEAR M F. Spoken dialogue technology: enabling the conversational user interface[J]. ACM Computing Surveys, 2002, 34(1): 90-169.
41. 赵新颜. 基于深度学习的对话系统研究与应用[D]. 合肥: 中国科学技术大学, 2022: 13-14.
42. NI J, YOUNG T, PANDELEA V, et al. Recent advances in deep learning based dialogue systems: a systematic survey[J]. Artificial Intelligence Review, 2022: 1-101.
43. JORGENSEN P E T, SONG M S, TIAN J. Operator theory, kernels, and feedforward neural networks[J]. arXiv preprint arXiv:2301.01327, 2023.
44. RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.
45. LI Z, LIU F, YANG W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 6999-7019
46. HEWAMALAGE H, BERGMEIR C, BANDARA K. Recurrent neural networks for time series forecasting: current status and future directions[J]. International Journal of Forecasting, 2021, 37(1): 388-427.
47. HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
48. YAO S, WAN X. Multimodal transformer for multimodal machine translation[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020: 4346-4350.
49. DU Y, PEI B, ZHAO X, et al. Deep scaled dot-product attention based domain adaptation model for biomedical question answering[J]. Methods, 2020, 173: 69-74.
50. MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, USA, 2017: 2227-2231.
51. CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition[C]. 41st IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 2016: 4960-4964.
52. KARMAKAR P, TENG S W, LU G. Thank you for attention: a survey on attention-based artificial neural networks for automatic speech recognition[J]. arXiv preprint arXiv:2102.07259, 2021.
53. WANG Y, SHEN Y, JIN H. A bi-model based RNN semantic frame parsing model for intent detection and slot filling[J]. arXiv preprint arXiv:1812.10235, 2018.
54. PHAM N Q, NGUYEN T S, NIEHUES J, et al. Very deep self-attention networks for end-to-end speech recognition[J]. arXiv preprint arXiv:1904.13377, 2019.
55. SYNNAEVE G, XU Q, KAHN J, et al. End-to-end ASR: from supervised to semi-supervised learning with modern architectures[J]. arXiv preprint arXiv:1911.08460, 2019.
56. LUO H, ZHANG S, LEI M, et al. Simplified self-attention for transformer-based end-to-end speech recognition[C]. 8th IEEE Spoken Language Technology Workshop, Shenzhen, China, 2021: 75-81.
57. ZHANG S, LIU C, JIANG H, et al. Feedforward sequential memory networks: a new structure to learn long-term dependency[J]. arXiv preprint arXiv:1512.08301, 2015.
58. MEHTA S, GHAZVININEJAD M, IYER S, et al. Delight: deep and light-weight transformer[J]. arXiv preprint arXiv:2008.00623, 2020.
59. WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: a survey[J]. arXiv preprint arXiv:2202.07125, 2022.
60. MISRA D. Mish: a self regularized non-monotonic activation function[J]. arXiv preprint arXiv:1908.08681, 2019.
61. LIN T, WANG Y, LIU X, et al. A survey of transformers[J]. AI Open, 2022, 1(1): 1-16.
62. WANG D, ZHANG X. Thchs-30: A free Chinese speech corpus[J]. arXiv preprint arXiv:1512.01882, 2015.
63. BU H, DU J, NA X, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline[C]. 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, Seoul, South Korea, 2017: 1-5.
64. ROUSSEAU A, DELÉGLISE P, ESTEVE Y. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks[C]. Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, 2014: 3935-3939.
65. WATANABE S, HORI T, KARITA S, et al. ESPNET: end-to-end speech processing toolkit[J]. arXiv preprint arXiv:1804.00015, 2018.
66. MALIK M, MALIK M K, MEHMOOD K, et al. Automatic speech recognition: a survey[J]. Multimedia Tools and Applications, 2021, 80: 9411-9457.
67. WHITEHEAD N, FIT-FLOREA A. Precision & performance: floating point and IEEE 754 compliance for NVIDIA GPUs[J]. RN (A+ B), 2011, 21(1): 18749-19424.
68. HIGUCHI Y, INAGUMA H, WATANABE S, et al. Improved mask-CTC for non-autoregressive end-to-end ASR[C]. 46th IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021: 8363-8367.
69. GHAHREMANI P, BABAALI B, POVEY D, et al. A pitch extraction algorithm tuned for automatic speech recognition[C]. 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014: 2494-2498.
70. KUDO T, RICHARDSON J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 2018: 66-71.
71. POVEY D, PEDDINTI V, GALVEZ D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]. 17th Annual Conference of the International Speech Communication Association, San Francisco, USA, 2016: 2751-2755.
72. SHAN C, WENG C, WANG G, et al. Component fusion: learning replaceable language model component for end-to-end speech recognition system[C]. 44th IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 2019: 5361-5635.
73. DONG L, XU S, XU B. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition[C]. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Canada, 2018: 5884-5888.
74. XU M, LI S, ZHANG X L. Transformer-based end-to-end speech recognition with local dense synthesizer attention[C]. 46th IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021: 5899-5903.
75. CHEN X, ZHANG S, SONG D, et al. Transformer with bidirectional decoder for speech recognition[J]. arXiv preprint arXiv:2008.04481, 2020.
76. PARK D S, CHAN W, ZHANG Y, et al. Specaugment: a simple data augmentation method for automatic speech recognition[J]. arXiv preprint arXiv:1904.08779, 2019.
77. KANNAN A, WU Y, NGUYEN P, et al. An analysis of incorporating an external language model into a sequence-to-sequence model[C]. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Canada, 2018: 1-5828.
78. WU T W, SU R, JUANG B. A label-aware BERT attention network for zero-shot multi-intent detection in spoken language understanding[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 4884-4896.
79. DEL PINO G E, GALAZ H. Statistical applications of the inverse gram matrix: a revisitation[J]. Brazilian Journal of Probability and Statistics, 1995: 177-196.
80. VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
81. QIN L, XU X, CHE W, ET AL. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling[J]. arXiv preprint arXiv:2004.10087, 2020.
82. HEMPHILL C T, GODFREY J J, DODDINGTON G R. The ATIS spoken language systems pilot corpus[C]. Speech and Natural Language: Proceedings of a Workshop, Hidden Valley, USA, 1990.
83. COUCKE A, SAADE A, BALL A, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces[J]. arXiv preprint arXiv:1805.10190, 2018.
84. MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
85. LIU B, LANE I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.
86. GANGADHARAIAH R, NARAYANASWAMY B. Joint multiple intent detection and slot labeling for goal-oriented dialog[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, USA, 2019: 564-569.
87. CHEN L, ZHOU P, ZOU Y. Joint multiple intent detection and slot filling via self-distillation[C]. 47th IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 2022: 7612-7616.
88. SÜZEN A A, DUMAN B, ŞEN B. Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-CNN[C]. 2nd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Online, 2020: 1-5.
89. LEE C H, CHENG H, OSTENDORF M. Dialogue state tracking with a language model using schema-driven prompting[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 4937-4949.
90. PENG B, LI X, GAO J, et al. Deep Dyna-Q: integrating planning for task-completion dialogue policy learning[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018: 2182-2192.
91. YU W, JIANG M, HU Z, et al. Knowledge-enriched natural language generation[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Online and Punta Cana, Dominican Republic, 2021: 11-16.
92. ŁAŃCUCKI A. Fastpitch: parallel text-to-speech with pitch prediction[C]. 46th IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 2021: 6588-6592.