Multimodal Understanding of Passenger Intents in Autonomous Vehicles

Presentation · December 2019

DOI: 10.13140/RG.2.2.24143.25762

4 authors, including:

Eda Okur
Intel
62 PUBLICATIONS **451** CITATIONS

SEE PROFILE

Multimodal Understanding of Passenger Intents in Autonomous Vehicles

Eda Okur
Intel Labs
Anticipatory Computing Lab
Hillsboro, OR 97124
eda.okur@intel.com

Shachi H. Kumar
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
shachi.h.kumar@intel.com

Saurav Sahay
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
saurav.sahay@intel.com

Lama Nachman
Intel Labs
Anticipatory Computing Lab
Santa Clara, CA 95054
lama.nachman@intel.com

1 Introduction

Understanding passenger intents from spoken interactions and car's vision (both inside and outside the vehicle) is an important building block towards developing contextual dialog systems for autonomous vehicles (AV). In this study, we continued exploring AMIE (Automated-vehicle Multimodal In-cabin Experience), the in-cabin agent responsible for handling multimodal passenger-vehicle interactions. When the passengers give instructions to AMIE, the agent should parse commands properly considering three modalities (i.e., verbal/language/text, vocal/audio, visual/video) and trigger the appropriate functionality of the AV system.

In previous work [6, 7], we collected a multimodal in-cabin dataset with multi-turn dialogues between the passengers and AMIE using a Wizard-of-Oz (WoZ) scheme, and we experimented with various RNN-based models to detect utterance-level intents (i.e., *set-destination*, *change-route*, *go-faster*, *go-slower*, *stop*, *park*, *pull-over*, *drop-off*, *open-door*, *other*) along with relevant slots associated with these intents.

In this work, we discuss the benefits of multimodal understanding of in-cabin utterances by incorporating verbal/language input together with the non-verbal/acoustic and visual input from inside and outside the vehicle. This ongoing research has potential impact of exploring real-world challenges with human-vehicle-scene interactions for autonomous driving support via spoken utterances.

2 Methodology

We explored leveraging multimodality for the Natural Language Understanding (NLU) module in the Spoken Dialogue System (SDS) pipeline. As our AMIE in-cabin dataset[1] has audio and video recordings, we investigated three modalities for the NLU: text, audio, and visual.

For text (verbal/language) modality, previous work [7] presents the details of our best-performing Hierarchical & Joint Bi-LSTM [11, 3, 15, 14] model (H-Joint-2)[2] and the results for utterance-level

[1] Details of AMIE data collection setup can be found in [12, 7].

[2] H-Joint-2: Detects/extracts *intent keywords & slots* using seq2seq Bi-LSTMs first (Level-1), then only the words that are predicted as *intent keywords & valid slots* are fed into Joint-2 model (Level-2), which is another seq2seq Bi-LSTM network for *utterance-level intent* detection (jointly trained with *slots* & *intent keywords*).

14th Women in Machine Learning Workshop (WiML 2019), co-located with the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

intent recognition and slot filling. These previous uni-modal results were obtained on the transcribed (i.e., via human transcriptions) and/or recognized (i.e., via Automatic Speech Recognition) text using GloVe word embeddings [9] as features.

In this study, we explore the following multimodal features to better assess in-cabin passenger intents[3] in autonomous vehicles:

2.1 Word and Speech Embeddings

We incorporated pre-trained speech embeddings, Speech2Vec[4], as features. These Speech2Vec embeddings are trained on a corpus of 500 hours of speech from LibriSpeech. We experimented with concatenating word and speech vectors using GloVe [9] (400K vocab, 100-dim), Speech2Vec [1] (37.6K vocab, 100-dim), and its Word2Vec [5] (37.6K vocab, 100-dim) counterpart. These Word2Vec embeddings are trained on the transcript of the same speech corpus (LibriSpeech).

2.2 Audio Features

Using openSMILE[5] [2], 1582 audio features are extracted for each utterance using the segmented audio clips from the in-cabin AMIE dataset. These features are the INTERSPEECH 2010 Paralinguistic Challenge (IS10) features including PCM loudness, MFCC, log Mel Freq. Band, LSP, etc. [10].

2.3 Visual Features

We extracted intermediate CNN features[6] from each video clip segmented per utterance in the AMIE dataset. Using the feature extraction process described in [4], visual descriptors are extracted from the activations of the intermediate convolution layers of a pre-trained CNN. We used the pre-trained Inception-ResNet-v2 model[7] [13] and generated 4096-dim features for each sample. We utilized two sources of visual information: (i) cabin/passenger view RGB video streams, (ii) road/outside view RGB camera recordings.

3 Experimental Results

Performance results of the compared intent recognition models with varying modality and feature concatenations can be found in Table 1, using hierarchical joint learning (H-Joint-2). We investigated incorporating the audio-visual features on top of text-only and text+speech embedding models. Using speech embeddings, as well as adding IS10 features from audio and intermediate CNN features from video brought improvements to our intent recognition models, reaching 0.92 F1-score.

Table 1: F1-scores of Intent Recognition with Multimodal Features (Embeddings & Audio & Visual)

Modalities	**Features**	**F1**(%)
Text	GloVe	89.02
Text & Audio	GloVe & Audio (openSMILE/IS10)	89.53
Text & Visual	GloVe & Video_cabin (CNN/Inception-ResNet)	89.40
Text & Visual	GloVe & Video_road (CNN/Inception-ResNet)	89.37
Text & Visual	GloVe & Video_cabin+road (CNN/Inception-ResNet)	89.68
Text & Audio	GloVe+Speech2Vec	90.85
Text & Audio	GloVe+Word2Vec+Speech2Vec	91.29
Text & Audio	GloVe+Word2Vec+Speech2Vec & Audio (IS10)	91.68
Text & Audio & Visual	GloVe+Word2Vec+Speech2Vec & Video_cabin (CNN)	91.50
Text & Audio & Visual	GloVe+Word2Vec+Speech2Vec & Video_cabin+road (CNN)	91.55

[3] Further details and in-cabin dataset statistics can be found in the short-paper version of this work [8].
[4] github.com/iamyuanchung/speech2vec-pretrained-vectors
[5] www.audeering.com/opensmile/
[6] github.com/MKLab-ITI/intermediate-cnn-features
[7] github.com/tensorflow/models/tree/master/research/slim

4 Conclusion

In this work, we briefly present our initial explorations towards multimodal understanding of passenger utterances in autonomous vehicles. We show that our experimental results outperformed the uni-modal text-only baseline results, and with multimodality, we achieved improved performances for passenger intent detection in AVs. These initial results may require further explorations for specific intents such as *stop* (e.g., audio intensity could have helped), or for relevant slots such as passenger *gesture/gaze* (e.g., cabin-view features) and outside *object*s (e.g., road-view features).

References

[1] Y.-A. Chung and J. Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Proc. INTERSPEECH 2018*, pages 811–815, 2018. doi: 10.21437/Interspeech.2018-2341. URL http://dx.doi.org/10.21437/Interspeech.2018-2341.

[2] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. ACM International Conference on Multimedia*, MM '13, pages 835–838, 2013. ISBN 978-1-4503-2404-5. doi: 10.1145/2502081.2502224. URL http://doi.acm.org/10.1145/2502081.2502224.

[3] D. Hakkani-Tur, G. Tur, A. Celikyilmaz, Y.-N. V. Chen, J. Gao, L. Deng, and Y.-Y. Wang. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. ISCA, June 2016. URL https://www.microsoft.com/en-us/research/publication/multijoint/.

[4] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International Conference on Multimedia Modeling*, pages 251–263. Springer, 2017. ISBN 978-3-319-51811-4. URL https://rd.springer.com/chapter/10.1007/978-3-319-51811-4_21.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

[6] E. Okur, S. H. Kumar, S. Sahay, A. A. Esme, and L. Nachman. Conversational intent understanding for passengers in autonomous vehicles. *13th Women in Machine Learning Workshop (WiML 2018), co-located with the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Dec 2018. URL http://arxiv.org/abs/1901.04899.

[7] E. Okur, S. H. Kumar, S. Sahay, A. A. Esme, and L. Nachman. Natural language interactions in autonomous vehicles: Intent detection and slot filling from passenger utterances. *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, April 2019. URL http://arxiv.org/abs/1904.10500.

[8] E. Okur, S. H. Kumar, S. Sahay, and L. Nachman. Towards multimodal understanding of passenger-vehicle interactions in autonomous vehicles: Intent/slot recognition utilizing audio-visual data. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2019)*, Sep 2019.

[9] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP'14)*, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[10] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010*, 2010. URL https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_2794.pdf.

[11] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, Nov. 1997. ISSN 1053-587X. doi: 10.1109/78.650093.

[12] J. Sherry, R. Beckwith, A. Arslan Esme, and C. Tanriover. Getting things done in an autonomous vehicle. In *Social Robots in the Wild Workshop, 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2018)*, March 2018. URL http://socialrobotsinthewild.org/wp-content/uploads/2018/02/HRI-SRW_2018_paper_3.pdf.

[13] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. URL http://arxiv.org/abs/1602.07261.

[14] L. Wen, X. Wang, Z. Dong, and H. Chen. Jointly modeling intent identification and slot filling with contextual and hierarchical information. In X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, editors, *Natural Language Processing and Chinese Computing*, pages 3–15, Cham, 2018. Springer. ISBN 978-3-319-73618-1. URL https://rd.springer.com/chapter/10.1007/978-3-319-73618-1_1.

[15] X. Zhang and H. Wang. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2993–2999, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3061040.