Towards Multimodal Understanding of Passenger-Vehicle Interactions in Autonomous Vehicles: Intent/Slot Recognition Utilizing Audio-Visual Data

Eda Okur **Shachi H Kumar** **Saurav Sahay** **Lama Nachman**
Intel Labs, Anticipatory Computing Lab, USA
{eda.okur, shachi.h.kumar, saurav.sahay, lama.nachman}@intel.com

1 Introduction

Understanding passenger intents from spoken interactions and car's vision (both inside and outside the vehicle) are important building blocks towards developing contextual dialog systems for natural interactions in autonomous vehicles (AV). In this study, we continued exploring AMIE (Automated-vehicle Multimodal In-cabin Experience), the in-cabin agent responsible for handling certain multimodal passenger-vehicle interactions. When the passengers give instructions to AMIE, the agent should parse such commands properly considering available three modalities (language/text, audio, video) and trigger the appropriate functionality of the AV system. We had collected a multimodal in-cabin dataset with multi-turn dialogues between the passengers and AMIE using a Wizard-of-Oz scheme via realistic scavenger hunt game.

In our previous explorations (Okur et al., 2018, 2019), we experimented with various RNN-based models to detect utterance-level intents (set destination, change route, go faster, go slower, stop, park, pull over, drop off, open door, and others) along with intent keywords and relevant slots (location, position/direction, object, gesture/gaze, time-guidance, person) associated with the action to be performed in our AV scenarios.

In this recent work, we propose to discuss the benefits of multimodal understanding of in-cabin utterances by incorporating verbal/language input (text and speech embeddings) together with the non-verbal/acoustic and visual input from inside and outside the vehicle (i.e., passenger gestures and gaze from in-cabin video stream, referred objects outside of the vehicle from the road view camera stream). Our experimental results outperformed text-only baselines and with multimodality, we achieved improved performances for utterance-level intent detection and slot filling.

2 Methodology

We explored leveraging multimodality for the NLU module in the SDS pipeline. As our AMIE in-cabin dataset[1] has video and audio recordings, we investigated 3 modalities for the NLU: text, audio, and video. For text (language) modality, our previous work (Okur et al., 2019) presents the details of our best-performing Hierarchical & Joint Bi-LSTM models (Schuster and Paliwal, 1997; Hakkani-Tur et al., 2016; Zhang and Wang, 2016; Wen et al., 2018) (H-Joint-2, see A) and the results for utterance-level intent recognition and word-level slot filling via transcribed and recognized (ASR output) textual data, using word embeddings (GloVe (Pennington et al., 2014)) as features. This study explores the following multimodal features:

Speech Embeddings: We incorporated pre-trained speech embeddings (Speech2Vec (Chung and Glass, 2018)) as features, trained on a corpus of 500 hours of speech from LibriSpeech. Speech2Vec[2] is considered as a speech version of Word2Vec (Mikolov et al., 2013) which is compared with Word2Vec vectors trained on the transcript of the same speech corpus. We experimented with concatenating word and speech embeddings by using pre-trained GloVe embeddings (6B tokens, 400K vocab, dim=100), Speech2Vec embeddings (37.6K vocab, dim=100), and its Word2Vec counterpart (37.6K vocab, dim=100).

Audio Features: Using openSMILE (Eyben et al., 2013), 1582 audio features are extracted for each utterance using the segmented audio clips from in-cabin AMIE dataset. These are the INTERSPEECH 2010 Paralinguistic Challenge features (IS10) including PCM loudness, MFCC, log Mel Freq. Band, LSP, etc. (Schuller et al., 2010).

[1] Details of AMIE data collection setup in (Sherry et al., 2018; Okur et al., 2019); in-cabin dataset statistics in A.

[2] github.com/iamyuanchung/speech2vec-pretrained-vectors

Modalities	Features (Embeddings)	Intent Recognition			Slot Filling		
		Prec	Rec	F1	Prec	Rec	F1
Text	GloVe (400K)	89.2	89.0	89.0	95.8	95.8	95.8
Text	Word2Vec (37.6K)	86.4	85.2	85.6	93.3	93.4	93.3
Audio	Speech2Vec (37.6K)	85.1	84.4	84.5	93.2	93.3	93.1
Text & Audio	Word2Vec + Speech2Vec	88.4	88.1	88.1	94.2	94.3	94.2
Text & Audio	GloVe + Speech2Vec	91.1	91.0	90.9	96.3	96.3	96.3
Text & Audio	GloVe + Word2Vec + Speech2Vec	91.5	91.2	91.3	96.6	96.6	96.6

Table 1: Speech Embeddings Experiments: Precision/Recall/F1-scores (%) of NLU Models

Modalities	Features	Prec	Rec	F1
Text	Embeddings (GloVe)	89.19	89.04	89.02
Text & Audio	Embeddings (GloVe) + Audio (openSMILE/IS10)	89.69	89.64	89.53
Text & Video	Embeddings (GloVe) + Video_cabin (CNN/Inception-ResNet-v2)	89.48	89.57	89.40
Text & Video	Embeddings (GloVe) + Video_road (CNN/Inception-ResNet-v2)	89.78	89.19	89.37
Text & Video	Embeddings (GloVe) + Video_cabin+road (CNN/Inception-ResNet-v2)	89.84	89.72	89.68
Text & Audio	Embeddings (GloVe+Word2Vec+Speech2Vec)	91.50	91.24	91.29
Text & Audio	Embeddings (GloVe+Word2Vec+Speech2Vec) + Audio (openSMILE)	91.83	91.62	91.68
Text & Audio & Video	Embeddings (GloVe+Word2Vec+Speech2Vec) + Video_cabin (CNN)	91.73	91.47	91.50
Text & Audio & Video	Embeddings (GloVe+Word2Vec+Speech2Vec) + Video_cabin+road (CNN)	91.73	91.54	91.55

Table 2: Multimodal (Audio & Video) Features Exploration: Precision/Recall/F1-scores (%) of Intent Recognition

Video Features: Using the feature extraction process described in (Kordopatis-Zilos et al., 2017), we extracted intermediate CNN features[3] for each segmented video clip from AMIE dataset. For any given input video clip (segmented for each utterance), one frame per second is sampled and its visual descriptor is extracted from the activations of the intermediate convolution layers of a pre-trained CNN. We used the pre-trained Inception-ResNet-v2 model[4] (Szegedy et al., 2016) and generated 4096-dim features for each sample. We experimented with adding 2 sources of visual information: (i) cabin/passenger view from the BackDriver RGB camera recordings, (ii) road/outside view from the DashCam RGB video streams.

3 Experimental Results

For incorporating speech embeddings experiments, performance results of NLU models on in-cabin data with various feature concatenations can be found in Table 1, using our previous hierarchical joint model (H-Joint-2). When used in isolation, Word2Vec and Speech2Vec achieves comparable performances, which cannot reach GloVe performance. This was expected as the pre-trained Speech2Vec vectors have lower vocabulary coverage than GloVe. Yet, we observed that concatenating GloVe + Speech2Vec, and further GloVe + Word2Vec + Speech2Vec yields better NLU results: F1-score increased from 0.89 to 0.91 for intent recognition, from 0.96 to 0.97 for slot filling.

[3]github.com/MKLab-ITI/intermediate-cnn-features
[4]github.com/tensorflow/models/tree/master/research/slim

For multimodal (audio & video) features exploration, performance results of the compared models with varying modality/feature concatenations can be found in Table 2. Since these audio/video features are extracted per utterance (on segmented audio & video clips), we experimented with the utterance-level intent recognition task only, using hierarchical joint learning (H-Joint-2). We investigated the audio-visual feature additions on top of text-only and text+speech embedding models. Adding openSMILE/IS10 features from audio, as well as incorporating intermediate CNN/Inception-ResNet-v2 features from video brought slight improvements to our intent models, reaching 0.92 F1-score. These initial results using feature concatenations may need further explorations, especially for certain intent-types such as stop (audio intensity) or relevant slots such as passenger gestures/gaze (from cabin video) and outside objects (from road video).

4 Conclusion

In this study, we present our initial explorations towards multimodal understanding of passenger utterances in autonomous vehicles. We briefly show that our experimental results outperformed certain baselines and with multimodality, we achieved improved overall F1-scores of 0.92 for utterance-level intent detection and 0.97 for word-level slot filling. This ongoing research has a potential impact of exploring real-world challenges with human-vehicle-scene interactions for autonomous driving support with spoken utterances.

References

Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Proc. INTERSPEECH 2018*, pages 811–815.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. ACM International Conference on Multimedia*, MM '13, pages 835–838.

Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Vivian Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. ISCA.

Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. 2017. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International Conference on Multimedia Modeling*, pages 251–263. Springer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. 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.

Eda Okur, Shachi H Kumar, Saurav Sahay, Asli Arslan Esme, and Lama Nachman. 2018. 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)*.

Eda Okur, Shachi H Kumar, Saurav Sahay, Asli Arslan Esme, and Lama Nachman. 2019. 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)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP'14)*.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010*.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

John Sherry, Richard Beckwith, Asli Arslan Esme, and Cagri Tanriover. 2018. 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)*.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.

Liyun Wen, Xiaojie Wang, Zhenjiang Dong, and Hong Chen. 2018. Jointly modeling intent identification and slot filling with contextual and hierarchical information. In *Natural Language Processing and Chinese Computing*, pages 3–15, Cham. Springer.

Xiaodong Zhang and Houfeng Wang. 2016. 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.

A Appendices

AMIE In-cabin Dataset: We obtained 1331 utterances having commands to AMIE agent from our in-cabin dataset. Annotation results for *utterance-level intent* types, *slots* and *intent keywords* can be found in Table 3 and Table 4.

AMIE Scenario	Intent Type	Utterance Count
Set/Change Destination/Route	SetDestination SetRoute	311 507
Finishing the Trip	Park PullOver Stop	151 34 27
Set/Change Driving Behavior/Speed	GoFaster GoSlower	73 41
Others (Door, Music, A/C, etc.)	OpenDoor Other	136 51
	Total	*1331*

Table 3: AMIE In-cabin Dataset Statistics: Intents

Slot/Keyword Type	Word Count
Intent Keyword	2007
Location	1969
Position/Direction	1131
Person	404
Time Guidance	246
Gesture/Gaze	167
Object	110
None	6512
Total	*12546*

Table 4: AMIE In-cabin Dataset Statistics: Slots

Hierarchical & Joint Model (H-Joint-2): 2-level hierarchical joint learning model that 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*) (Okur et al., 2019).