Conversational Intent Understanding for Passengers 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

Asli Arslan Esme
Intel Labs
Anticipatory Computing Lab
Hillsboro, OR 97124
asli.arslan.esme@intel.com

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

1 Introduction

Understanding passenger intents and extracting relevant slots are important building blocks towards developing a contextual dialogue system responsible for handling certain vehicle-passenger interactions in autonomous vehicles (AV). When the passengers give instructions to AMIE (Automated-vehicle Multimodal In-cabin Experience), the agent should parse such commands properly and trigger the appropriate functionality of the AV system. In our AMIE scenarios, we describe usages and support various natural commands for interacting with the vehicle. We collected a multimodal in-cabin data-set with multi-turn dialogues between the passengers and AMIE using a Wizard-of-Oz scheme. We explored various recent Recurrent Neural Networks (RNN) based techniques and built our own hierarchical models to recognize passenger intents along with relevant slots associated with the action to be performed in AV scenarios. Our experimental results achieved F1-score of 0.91 on utterance-level intent recognition and 0.96 on slot extraction models.

2 Methodology

Our AV in-cabin data-set includes 30 hours of multimodal data collected from 30 passengers (15 female, 15 male) in 20 rides/sessions. 10 types of passenger intents are identified and annotated as: *Set/Change Destination*, *Set/Change Route* (including turn-by-turn instructions), *Go Faster*, *Go Slower*, *Stop*, *Park*, *Pull Over*, *Drop Off*, *Open Door*, and *Other* (turn music/radio on/off, open/close window/trunk, change AC/temp, show map, etc.). Relevant slots are identified and annotated as: *Location*, *Position/Direction*, *Object*, *Time-Guidance*, *Person*, *Gesture/Gaze* (this, that, over there, etc.), and *None*. In addition to utterance-level intent types and their slots, word-level intent keywords are annotated as *Intent* as well. We obtained 1260 unique utterances having commands to AMIE from our in-cabin data-set. We expanded this data-set via Amazon Mechanical Turk and ended up with 3347 utterances having intents. The annotations for intents and slots are obtained on the transcribed utterances by majority voting of 3 annotators.

For slot filling and intent keywords extraction tasks, we experimented with seq2seq LSTMs and GRUs, and also Bidirectional LSTM/GRUs. The passenger utterance is fed into a Bi-LSTM network via an embedding layer as a sequence of words, which are transformed into word vectors. We also experimented with GloVe, word2vec, and fastText as pre-trained word embeddings. To prevent

13th Women in Machine Learning Workshop (WiML 2018), co-located with the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Table 1: Slot Extraction Results (10-fold CV)

Slot Type	**F1**
Location	0.95
Position	0.95
Person	0.97
Object	0.89
Time Guidance	0.85
Gesture	0.92
None	0.98
AVERAGE	***0.96***

Table 2: Intent Keyword Extraction Results (10-fold CV)

Keyword Type	**F1**
Intent	0.94
Non-Intent	0.99
AVERAGE	***0.98***

overfitting, a dropout layer is used for regularization. Best performing results are obtained with Bi-LSTMs and GloVe embeddings (6B tokens, 400K vocab size, dim 100).

For utterance-level intent detection, we experimented with mainly 5 models: (1) Hybrid: RNN + Rule-based, (2) Separate: Seq2one Bi-LSTM + Attention, (3) Joint: Seq2seq Bi-LSTM for slots/intent keywords & utterance-level intents, (4) Hierarchical + Separate, (5) Hierarchical + Joint. For (1), we extract intent keywords/slots (Bi-LSTM) and map them into utterance-level intent types (rule-based via term frequencies for each intent). For (2), we feed the whole utterance as input sequence and intent-type as single target. For (3), we experiment with the joint learning models [1, 2, 5] where we jointly train word-level intent keywords/slots and utterance-level intents (adding <BOU>/<EOU> terms to the start/end of utterances with intent types). For (4) and (5), we experiment with the hierarchical models [6, 3, 4] where we extract intent keywords/slots first, and then only feed the predicted keywords/slots as a sequence into (2) and (3), respectively.

3 Experimental Results

The slot extraction and intent keywords extraction results are given in Table 1 and Table 2, respectively. Table 3 summarizes the results of various approaches we investigated for utterance-level intent understanding. Table 4 shows the intent-wise detection results for our AMIE scenarios with the best performing utterance-level intent recognizer.

Table 3: Utterance-level Intent Recognition Results (10-fold CV)

Utterance-level Intent Detection Models	**F1**
Hybrid-1: RNN + Rule-based (*intent keywords*)	0.86
Hybrid-2: RNN + Rule-based (*intent keywords & slots*)	0.90
Separate-1: Seq2one Bi-LSTM	0.88
Separate-2: Seq2one Bi-LSTM + Attention (withContext)	0.89
Joint: Seq2seq Bi-LSTM (*intent keywords & slots & utterance-level intent types*)	0.88
Hierarchical & Separate-1	0.90
Hierarchical & Separate-2 (Separate-1 + Attention)	0.90
Hierarchical & Joint	**0.91**

Table 4: Intent-wise Performance Results of Utterance-level Intent Recognition Models: Hierarchical & Joint (10-fold CV)

AMIE Scenario	Intent Type	F1
Finishing the Trip Use-cases	Stop	0.92
	Park	0.94
	PullOver	0.96
	DropOff	0.95
Set/Change Destination/Route	Set/ChangeDest	0.89
	Set/ChangeRoute	0.88
Set/Change Driving Behavior/Speed	GoFaster	0.88
	GoSlower	0.88
Others (Door, Music, A/C, etc.)	OpenDoor	0.95
	Other	0.82
	AVERAGE	***0.91***

4 Conclusion

After exploring various recent Recurrent Neural Networks (RNN) based techniques, we built our own hierarchical joint models to recognize passenger intents along with relevant slots associated with the action to be performed in AV scenarios. Our experimental results outperformed certain competitive baselines and achieved overall F1-scores of 0.91 for utterance-level intent recognition and 0.96 for slot extraction tasks.

References

[1] D. Hakkani-Tür, 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/.

[2] B. Liu and I. Lane. Joint online spoken language understanding and language modeling with recurrent neural networks. *CoRR*, abs/1609.01462, 2016. URL http://arxiv.org/abs/1609.01462.

[3] Z. Meng, L. Mou, and Z. Jin. Hierarchical rnn with static sentence-level attention for text-based speaker change detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2203–2206, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4918-5. doi: 10.1145/3132847.3133110. URL http://doi.acm.org/10.1145/3132847.3133110.

[4] 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 International Publishing. ISBN 978-3-319-73618-1.

[5] 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. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3061040.

[6] Q. Zhou, L. Wen, X. Wang, L. Ma, and Y. Wang. A hierarchical lstm model for joint tasks. In M. Sun, X. Huang, H. Lin, Z. Liu, and Y. Liu, editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 324–335, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47674-2.