Improving Intent Classification in an E-commerce Voice Assistant by Using Inter-Utterance Context

Arpit Sharma
@WalmartLabs, Sunnyvale, USA
arpit.sharma@walmartlabs.com

Abstract

In this work, we improve the intent classification in an English based e-commerce voice assistant by using inter-utterance context. For increased user adaptation and hence being more profitable, an e-commerce voice assistant is desired to understand the context of a conversation and not have the users repeat it in every utterance. For example, let a user's first utterance be *'find apples'*. Then, the user may say *'i want organic only'* to ***filter*** out the results generated by an assistant with respect to the first query. So, it is important for the assistant to take into account the context from the user's first utterance to understand her intention in the second one. In this paper, we present our approach for contextual intent classification in Walmart's e-commerce voice assistant. It uses the intent of the previous user utterance to predict the intent of her current utterance. With the help of experiments performed on real user queries we show that our approach improves the intent classification in the assistant.

1 Introduction

Recently, there has been a notable advancement in the field of voice assistants[1]. Consequently, voice assistants are being deployed heavily in lucrative domains such as e-commerce (Mari et al., 2020), customer service (Cui et al., 2017) and healthcare (Mavropoulos et al., 2019). There are various e-commerce based voice assistants available in the market, including Amazon's Alexa voice shopping (Maarek, 2018) and Walmart's on Google assistant/Siri [2]. Their goal is to free us from the tedious task of buying stuff by visiting stores and websites. A major challenge in the fulfilment of this goal is their capability to precisely understand an utterance in a dialog without providing much context in the utterance. For example, an assistant must precisely understand that when a user says *'five'* after a query to *add bananas* to her cart then she intents to ***add*** five bananas to her cart. Whereas if a user says *'five'* as her first utterance to the shopping assistant then her intention is ***unknown*** (i.e., it does not represent any e-commerce action at the start of a conversation). Handling such scenarios require the Natural Language Understanding (NLU) component in the assistants to utilize the context while predicting the intent associated with an utterance. The current intent prediction systems (Chen et al., 2019; Goo et al., 2018; Liu and Lane, 2016) do not focus on such contextual dependence. In this work we integrate inter-utterance contextual features in the NLU component of the shopping assistant of Walmart company to improve its intent classification.

There are four main aspects of a voice assistant, namely, Speech-to-Text, NLU, Dialog Management (DM) and Text-to-Speech. The NLU component (such as the one in Liu and Lane (2016)) identifies intent(s) and entities in a user utterance. The dialog manager uses the output of the NLU component to prepare a suitable response for the user. The NLU systems in the currently available voice enabled shopping assistants[3] do not focus on inter-utterance context and hence the onus of context disambiguation lies upon the dialog manager. Although it is possible to capture a small number of such cases in the dialog manager, it becomes difficult for it to scale for large number of contextually dependent utterances. For example, let us consider the user utterance *'five'* after the utterance to add something to her cart. Then a dialog manager can predict its intent by using the rule: *if previous_intent =* ***add_to_cart*** *and the query is an* ***integer*** *then intent =* ***add_to_cart*** *else intent =* ***un-***

[1] https://bit.ly/2Xr71xv
[2] https://bit.ly/33TmwkN, https://bit.ly/2JqH9eO
[3] https://bit.ly/2ZAa5KA

Proceedings of the 3rd Workshop on e-Commerce and NLP (ECNLP 3), pages 40–45
Online, July 10, 2020. ©2020 Association for Computational Linguistics

known. But such a general rule can not be created for many other queries such as *'organic please'* (previous intent = *add_to_cart*, intent = *filter*) and *'stop please'* (previous intent = *add_to_cart*, intent = *stop*).

In this paper, we present our work to improve the intent classification in the shopping assistant of Walmart company by using inter-utterance context. Our work also reduces the contextual disambiguation burden from the dialog manager. Here, we implement our approach using two neural network based architectures. With the help of experiments we also compare the two implementations.

2 Related Work

Various intent classification works (Chen et al., 2019; Goo et al., 2018; Liu and Lane, 2016) have been proposed in the recent years. Most of them mainly focus on the current utterance only and try to predict its intent based on the information present in it. We take it one step further and use the context from the immediately previous utterance.

A work which focuses on the contextual information while predicting the intents is mentioned in Naik et al. (2018). It uses visual information as context. Although it is useful in an assistant which comes with a screen, it is not applicable for a voice only assistant.

Another work (Mensio et al., 2018) uses an entire conversation (intents of all the utterances and all the replies from the assistant) as context. Although it makes sense to use the entire conversation as context in a general purpose chat-bot, we believe that in the e-commerce domain a conversation between an assistant and a user is fragmented into smaller goals such as *'finding an item in the inventory'* and *'adding an item in cart'*. Each such goal should be defined by a small number of utterances only. This is because the objective of the voice assistant is to simplify the shopping experience for the user instead of engaging her in a long conversation for a simpler task. So, in our work we use the intent of the previous utterance only as the context for the intent prediction of the current utterance.

3 Data Generation & Our Approach

In this section we provide the data generation details and a detailed overview of our context aware intent classification implementations.

3.1 Contextual Data Generation

In this work our main goal is to improve the intent classification for e-commerce related utterances. We achieve this goal by using inter-utterance context. There are various data sets (ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018)) for intent classification. None of them contain contextual data instances. Furthermore, they do not focus on e-commerce related data. So, as part of this work we generated a context aware data set for e-commerce specific queries. We used a template based approach to generate the data. It was inspired by our previous work as mentioned in Sharma et al. (2018). Following are the two main steps in the data generation phase.

- **Step 1:** A set of 1735 templates (corresponding to 32 intents) are curated over a period of time by product managers and data scientists. An example of a template is "`i wanted [brand]`" where `[brand]` is a placeholder for a product's brand (such as *Shamrock Farms*). The templates are used to generate template and intent combinations. Each such combination contains a previous intent, a template and the template's correct intent. For example, following are two of the combinations generated with respect to the above mentioned template, 1)
  1. previous intent = `search`, template = `i wanted [brand]`, intent = `filter`,
  2. previous intent = `START`, template = `i wanted [brand]`, intent = `unknown`.
- **Step 2:** In this step the placeholders in the template and intent combinations are replaced with their possible values from predefined sets to generate the data points. Multiple data points are generated from each template combination by using the placeholders' values (10 distinct values for each placeholder in each template). The possible values of the placeholders are extracted from the products' catalog of Walmart company.

3.1.1 Fixing Unbalanced Data

The data generated by using the above steps was found to be unbalanced in the following two ways.

1. We found that the generated data contained more instances corresponding to some intents than

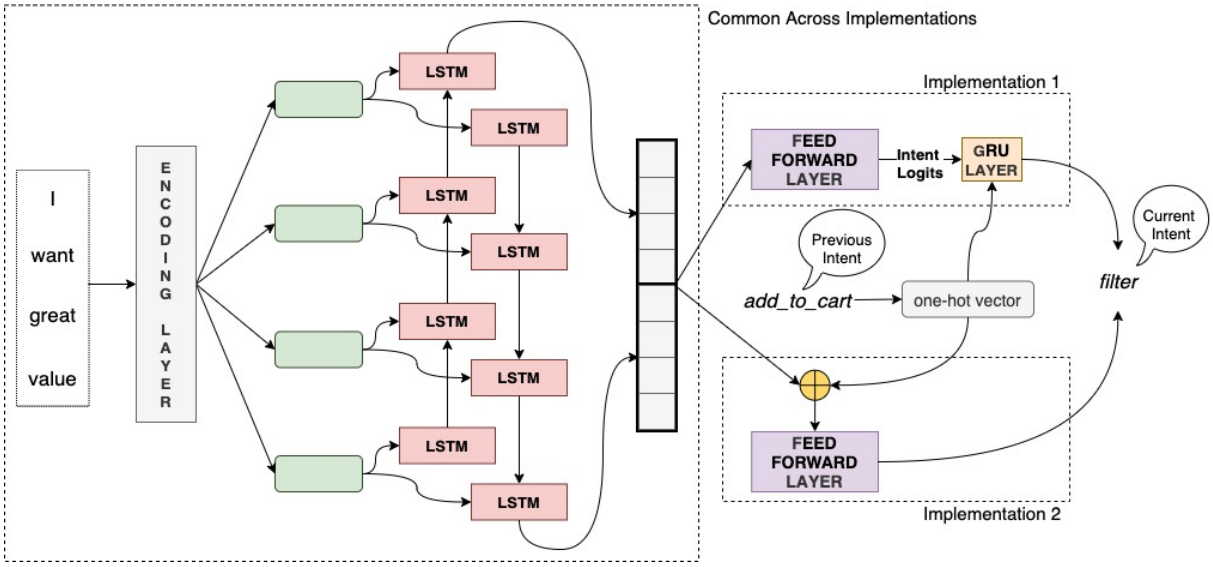

Figure 1: Deep Learning Architectures of the Implementations.
Implementation 1 is Bi-LSTM+FeedForward+GRU and Implementation 2 is Bi-LSTM+FeedForward

others. This is because the templates corresponding to some intents (such as *add to cart*) contained placeholders which were replaced by many possible values whereas the templates corresponding to other intents (such as *stopping a conversation*) did not contain any placeholders. To resolve this kind of skewness we performed Random Oversampling (ROS) (Rathpisey and Adji, 2019) with respect to the intents with minimal data.

2. The contextual instances in the generated data were overwhelmed by their non-contextual parts. Let `T` be a contextual template such that its intent is `I` iff the previous intent is `P` otherwise its intent is `unknown`. Since there are a total of 32 intents, `T` corresponds to 32 contextual template combinations of previous intent and current intent. In 31 among those, the current intent of `T` is `unknown` whereas only one corresponds to `I`. To resolve this kind of imbalance we performed Random Oversampling (ROS) (Rathpisey and Adji, 2019) with respect to the one combination mentioned above.

3.2 Our Approach

We used two different neural network architectures for two separate implementations of our approach. The two architectures are as shown in the Figure 1. Following are the details of the architectures.

1. **Bi-LSTM+FeedForward+GRU**: The first architecture is inspired by the work in Mensio et al. (2018). It has two main components. First, a bi-LSTM (Huang et al., 2015; Plank et al., 2016) encoder which generates a vector encoding of an input utterance. The vector represents a summarized version of the utterance. It is generated from the embeddings of the words in the query. We experimented with different pre-trained language models (BERT (Devlin et al., 2018) and Glove (Pennington et al., 2014)) to retrieve the initial word embedding. Second component is a GRU (Chung et al., 2014) layer whose inputs consist of the output of a feed forward layer with respect to the embedding generated by the bi-LSTM encoder, along with the one-hot encoding of the previous utterance's intent. The output of this layer is the intent of the input utterance (See Figure 1).

2. **Bi-LSTM+FeedForward**: Similar to the first architecture, the second one also has two components, a bi-LSTM layer followed by a feed forward layer. As in the first architecture, the bi-LSTM layer generates a vector which represents a user utterance. The output of the bi-LSTM layer is then concatenated with the one-hot encoding of the previous intent and entered as an input to a feed forward layer (See Figure 1).

4 Evaluation & Results

The main goal of this work is to improve the intent classification in the NLU component of Walmart's shopping assistant by using inter-utterance context. In this section we present the quantitative details of the data, the experiments and the analysis of the results.

Implementation	Oversampling	Intent Accuracy on User Logs (% correct)		
		Glove 6B	Glove 840B	BERT
Bi-LSTM+FeedForward+GRU	✔	83.05	80.51	**87.68**
	✖	79.12	73.64	72.98
Bi-LSTM+FeedForward	✔	86.5	86.67	86.34
	✖	82.58	83.39	70.57

Table 1: Evaluation results for Contextual Intent Classification on 2550 Real User Queries

4.1 Dataset Details

We followed the data generation steps mentioned in the Section 3.1 with and without oversampling to generate a set of 730K and 570K instances respectively. Each set is split into training (85%) and validation (15%). For testing the trained models we used the real user queries which were taken from the live logs of Walmart's shopping assistant. We selected all the queries from two weeks of live logs. Then we filtered the retrieved queries by keeping only the unique ones. By following the above steps, we got 2550 unique user queries in our test set.

4.2 Experiment 1

The hypothesis that led to this work states that contextual evidence is needed to find the correct intent of a user utterance in an e-commerce voice assistant. To test the hypothesis, in this experiment we analyzed the set of 2550 unique user utterances from user logs. 40.11% of those (i.e., 1023) were found to be contextual. For example *'give me a smaller size'* is one such contextual query which when appears after an *add to cart* utterance, implies *filtering* the results of the previous utterance whereas when appeared first in a conversation between a user and a voice assistant, does not make sense (or *unknown* intent). The *add to cart* and *search* intents are most popular among the logs. We found that 1005 out of 1023 contextual queries were related to those intents. This emphasizes the importance of contextual disambiguation even further. 917 (approx. 90%) among 1023 were correctly classified by the best performing version (BERT based) of our implementations (see Table 1). The current, non-contextual intent classifier of the Walmart's shopping assistant classifies all the contextual queries as one intent (say *abc*). The contextual disambiguation burden lies on the dialog manager by using rules (as mentioned in the Section 1). Presently, a total of 12 rules exist to handle contextual templates corresponding to one intent (out of 32) only. We also found that out of 1023 contextual queries, about 88% are classified by the non-contextual intent classifier as *abc*.

4.3 Experiment 2

In this experiment, we tested our implementations with respect to different input word embedding and data (with oversampling and without oversampling). We used BERT's huggingface[4] pretrained embedding (length=768), and Glove[5] 6 billion and 840 billion pre-trained embedding (length=300). The evaluation results are as shown in the Table 1. Each model was trained for 10 epochs. The BERT embedding of a word was calculated by taking an average of the last 4 layers in the 12-layer BERT pre-trained model[5].

4.4 Results Analysis

The results of experiment 1 show the usefulness of contextual intent classification.

The results of experiment 2 show that the Bi-LSTM and GRU based implementation performs best with an overall accuracy of 87.68% on all the live user logs and approximately 90% on the contextual logs.

Inference speed plays an important role in production deployment of models. Although the performance of BERT based Bi-LSTM+FeedForward+GRU is better than the Glove (840B) based Bi-LSTM+FeedForward, the latency of first (450 milliseconds, averaged over 2550 queries on CPU) one is considerably more than the second (5 milliseconds on CPU). See Table 2 for inference speeds of the different models.

We observed that many (about 50%) errors in both the implementations were caused by the inability of the training data to represent real user data. A way to address such errors is by using real user queries also to train the models. It requires manual effort to label user logs. We are currently in the process of such an effort through crowd workers.

[4]https://github.com/huggingface/transformers
[5]https://nlp.stanford.edu/projects/glove/

Implementation	Word Embedding	Inference Speed (CPU)
Bi-LSTM+FeedForward+GRU	Glove 6B	$\approx$ 5 ms
Bi-LSTM+FeedForward+GRU	Glove 840B	$\approx$ 5 ms
Bi-LSTM+FeedForward+GRU	BERT	$\approx$ 450 ms
Bi-LSTM+FeedForward	Glove 6B	$\approx$ 5 ms
Bi-LSTM+FeedForward	Glove 840B	$\approx$ 5 ms
Bi-LSTM+FeedForward	BERT	$\approx$ 450 ms

Table 2: Inference Speed of Different Models (Averaged Over 2550 Real User Queries)

We believe that using a combination of templates based and real user data will improve the accuracy of the implementations even further.

5 Conclusion & Future Work

In this paper, we presented our work of improving intent classification in Walmart's shopping assistant. We used previous utterance's intent as context to identify current utterance's intent. The contextual update in the NLU layer (intent classification) also takes the burden of intent based contextual disambiguation away from a dialog manager. As hypothesized, the experimental results show that our approach improves the intent classification by handling contextual queries. We presented two implementations of our approach and compared them with respect to live user logs.

Though, in this work our main focus was on the contextual disambiguation of intents, the entities are also contextually dependent. For example *'five'* uttered after *'add bananas'* may refer to the quantity *five* whereas if uttered after *'pick a delivery time'* may refer to the time of day *five (am/pm)*. In future we would like to use contextual features to disambiguate between entities and improve the entity tagging as well.

Acknowledgments

I would like to acknowledge the support of my colleagues in the Emerging Technologies team at WalmartLabs, Sunnyvale, CA office. They provided their valuable feedback which helped in maturing this submission to a publishable states.

References

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, pages 97–102.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Yoelle Maarek. 2018. Alexa and her shopping journey. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1–1.

Alex Mari, Andreina Mandelli, and René Algesheimer. 2020. The evolution of marketing in the context

of voice commerce: A managerial perspective. In *Proceeding of the 22nd International Conference on Human-Computer Interaction*.

Thanassis Mavropoulos, Georgios Meditskos, Spyridon Symeonidis, Eleni Kamateri, Maria Rousi, Dimitris Tzimikas, Lefteris Papageorgiou, Christos Eleftheriadis, George Adamopoulos, Stefanos Vrochidis, et al. 2019. A context-aware conversational agent in the rehabilitation domain. *Future Internet*, 11(11):231.

Martino Mensio, Giuseppe Rizzo, and Maurizio Morisio. 2018. Multi-turn qa: A rnn contextual approach to intent classification for goal-oriented systems. In *Companion Proceedings of the The Web Conference 2018*, pages 1075–1080.

Vishal Ishwar Naik, Angeliki Metallinou, and Rahul Goel. 2018. Context aware conversational understanding for intelligent agents with a screen. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529*.

Heng Rathpisey and Teguh Bharata Adji. 2019. Handling imbalance issue in hate speech classification using sampling-based methods. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 193–198. IEEE.

Arpit Sharma, Vivek Kaul, and Shankara B Subramanya. 2018. Semantic representation and parsing for voice ecommerce. In *R2K Workshop: Integrating learning of Representations and models with deductive Reasoning that leverages Knowledge (Co-located with KR 2018)*.