IMPROVING DEVICE DIRECTEDNESS CLASSIFICATION OF UTTERANCES WITH SEMANTIC LEXICAL FEATURES

Kellen Gillespie[†], *Ioannis C. Konstantakopoulos*[†], *Xingzhi Guo*[⋆], *Vishal Thanvantri Vasudevan*[†], *Abhinav Sethy*[†]

[†] Amazon.com, Inc.
[⋆] Stony Brook University

ABSTRACT

User interactions with personal assistants like Alexa, Google Home and Siri are typically initiated by a wake term or wakeword. Several personal assistants feature "follow-up" modes that allow users to make additional interactions without the need of a wakeword. For the system to only respond when appropriate, and to ignore speech not intended for it, utterances must be classified as device-directed or non-device-directed. State of the art systems have largely used acoustic features for this task, while others have used only lexical features or have added LM-based lexical features. We propose a directedness classifier that combines semantic lexical features with a lightweight acoustic feature and show it is effective in classifying directedness. The mixed-domain lexical and acoustic feature model is able to achieve 14% relative reduction of EER over a state of the art acoustic-only baseline model. Finally, we successfully apply transfer learning and semi-supervised learning to the model to improve accuracy even further.

Index Terms— Directedness, Semantic Classification, LSTM, Semi-supervised Learning, Word Embeddings

1. INTRODUCTION

Personal assistants such as Alexa, Google Home, and Siri are becoming commonplace in modern households [1]. These systems respond to user commands prefixed by a wake command or *wakeword*, e.g. *"Alexa, what time is it?"* or *"Hey Google, what's the weather?"*. This wakeword acts as signal that the utterance is intended for the assistant, in other words *device-directed*.

To make consecutive commands more natural and conversational, personal assistants have added "follow-up" modes that allow users to give consecutive commands without repeated use of the wakeword. Recent examples include Baidu DuerOS's "Full-Duplex" feature, Google Home's "Continued Conversation" feature and Amazon Alexa's "Follow-up Mode" [2, 3, 4]. These features allow for more natural conversations as shown in Table 1.

STANDARD MODE	FOLLOW-UP MODE
COMPUTER, WHAT'S THE WEATHER? (DEVICE RESPONDS)	COMPUTER, WHAT'S THE WEATHER? (DEVICE RESPONDS)
COMPUTER, WHAT ABOUT TOMORROW? (DEVICE RESPONDS)	WHAT ABOUT TOMORROW? (DEVICE RESPONDS)
COMPUTER, THANK YOU	THANK YOU

Table 1. More Natural Conversation with Follow-up Modes

To achieve this behavior, the device's microphone is automatically re-opened after the device responds to the user. Users can then give device-directed follow-up commands without the need of a wakeword. However, if a user has no follow-up commands to give, the system could pick up speech not intended for it or *non-device-directed* speech. Non-device-directed speech includes background speech, speech from media sources, and any other utterance not intended for the device. To alleviate this problem, and to avoid responding to non-device-directed utterances, virtual assistants employ directedness classifiers to separate utterances intended for the device from those that are not.

Directedness classification is an important task in many systems. Many such systems use acoustic features to determine directedness. In the personal assistant space, Mallidi et al. [5] leverage ASR decoder features as well as acoustic embeddings from the audio signal to perform directedness classification for one of the aforementioned follow-up modes. We use this work as the acoustic-only baseline in our evaluations.

Another system in the personal assistant space makes use of acoustic features and improves performance through the addition of an attention mechanism [6]. In related smart-environment applications, Reich et al. [7] use ASR decoder and prosodic features to perform directedness classification and Paek et al. [8] combine signal and user behavior characteristics to determine directedness in a continuous listening presentation domain.

Other systems rely on language features to determine directedness [9]. In a simulated planetary exploration scenario, Dowding, et al. [10] develop a system to help robotic members of mixed human robot teams determine if utterances were intended for them or other members. Shriberg et al. [11] add LM-based lexical features and contextual similarity fea-

STRUCTURED	UNSTRUCTURED
DID YOU REORDER YOUR PILLS	IT'S IT'S JUST FLASHING YELLOW
I DON'T KNOW WHAT SHE JUST ORDER	CAN YOU HEY COMPUTER
WELL YOU GOT FOUR MORE HOURS	BREAK AT A BIGGER
MOM WHAT DID YOU SAY	TELL US MOVING
WHAT ARE YOU DOING	WEATHER TALKING ABOUT HAL

Table 2. Structured and Unstructured Utterances Misidentified as Device-directed

tures to supplement acoustic features in a Human-Human-Computer dialog environment.

Our proposed model will leverage signals from both acoustic and lexical domains. Other systems that have combined these domains have used LM-based or character-level lexical features [5, 11]. We propose a system to combine acoustic features with word-level semantic lexical features to improve directedness classification. We further improve the model with contextual features and attention mechanisms [6, 11]. To reduce the data labeling burden required by the model, we use transfer learning from a related dataset and experiment with semi-supervised learning techniques.

The paper is organized as follows. In Section 2, we look at the follow-up dataset to motivate the features and architectures to be evaluated. In Section 3, we describe our baseline feed-forward model as well as our proposed LSTM architectures. In Section 4, we evaluate the discussed models and features and discuss the results. In Section 5, we examine the effectiveness of a simple semi-supervised learning technique on our model training. In Section 6, we discuss our overall findings and plans to improve our model.

2. DATA ANALYSIS

The primary feature of our proposed model is token-level semantic embeddings of the current turn utterance text derived from pre-trained word embeddings. Sentence embeddings can typically be derived by averaging token-level embeddings, but in this section we examine the data to identify additional features and architectures to develop a better model.

2.1. Utterance-level Structure

Many non-device-directed utterances in our data, due to factors such as containing multiple speakers or ASR errors, show a lack of sentence structure. Table 2 shows some examples of structured and non-structured utterances from the non-device-directed class.

A model that accounts for word order, and thus indirectly models grammatical structure, can be expected to learn to reject such unstructured utterances as non-device-directed. For this reason, we expect a recurrent model such as an LSTM will outperform a baseline DNN model on this task.

2.2. Contextually-Relevant Utterances

While the current turn utterance can normally provide most of the information needed to perform directedness classification, there is a subset of our dataset that contains contextually-relevant utterances. We give a few examples in Table 3.

PREVIOUS TURN	CURRENT TURN
COMPUTER ADD BANANAS TO MY SHOPPING LIST	ADD CHICKEN
COMPUTER WHAT'S THE WEATHER FOR TODAY	AND FOR TOMORROW
COMPUTER WHO IS ADELE?	PLAY
COMPUTER ARE YOU UNIQUE?	ARE YOU ONE OF A KIND?

Table 3. Utterances with Relevant Previous Turns

Looking solely at the current turn, these examples may be difficult to classify. With the context of the previous turn, however, they more clearly become interpretable as device-directed utterances. For this reason, we experiment with incorporating the previous turn text into the features to aid in classification.

2.3. Textual Ambiguity

While we expect semantic lexical features to aid in directedness classification, they do not on their own solve the problem of textual ambiguity present in our data. Many utterances belong to both classes in our dataset, in fact 29.7% of the dataset consists of utterances belonging to both classes. We highlight some examples below in Table 4, showing the most frequent occurring utterances belonging to both classes. For each utterance, we look at class frequencies to determine prior probabilities of each belonging to each class.

UTTERANCE	P(DD)	P(NDD)
THANK YOU	93.2	6.8
STOP	44.8	55.2
OKAY	45.8	55.2
CANCEL	70.7	29.3
WHAT	32.6	67.4
NEXT	62.8	37.2
GOOD NIGHT	61.6	38.4
PLAY	36.8	63.2

Table 4. Textual ambiguity of utterances common to device-directed (DD) and non-device-directed (NDD) classes.

It's clear that for many frequent utterances, text-only input will be insufficient for distinguishing between classes. For this reason, we will fuse the lexical features with acoustic features to assist in separating these ambiguous cases. Our acoustic feature will be token-level confidence scores from an upstream ASR decoding component [12]. We also expect a subset of these cases to be assisted by the contextual features described in Section 2.2.
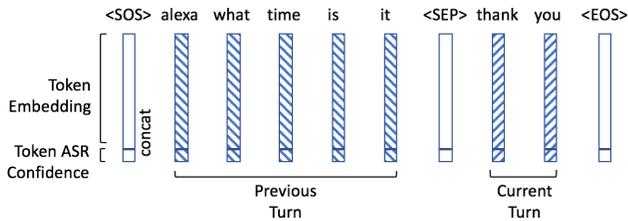
Fig. 1. Input Feature Representations

3. CANDIDATE MODELS

3.1. Feature Representation

We experiment with three features for our candidate models: the ASR-produced text of the current turn, the text of the previous turn, and token-level ASR confidence scores. The turns are separated by special tokens similar to those used in other systems [13]. We concatenate the lexical features, the word embeddings of each token, with their corresponding acoustic feature in the form of the ASR confidence scores. For pre-trained embeddings we use fastText wiki news word vectors [14] with an embedding size of 300. The only token embedding we specifically train for our data is an out-of-vocabulary or OOV token to better represent the OOV tendencies of personal assistant commands, while all other embeddings are left as-is. Due to the trained OOV token we do not make use of subword vectors in fastText. Figure 1 shows the overall input representation of our features.

3.2. AVG-DNN Model

The non-recurrent proposed model is a fully-connected feed-forward network using a word average sentence representation of the utterance as input. The sentence embedding is simply the average of the token features shown in Figure 1. The DNN has 3 layers: an input layer, dense hidden layer of size 150, and an output layer.

3.3. LSTM Model

The LSTM model [15] makes use of sequential information in the input features. For this reason, the token-level features in Figure 1 are used as-is, with each token representing an input frame. The LSTM has 3 layers of size 150 followed by a single dense layer and an output layer. Bidirectional LSTM (BLSTM) models were tested but did not improve over the single direction versions.

3.4. LSTM+Attention Model

Attention mechanisms are a common addition to LSTM architectures in neural translation and even directedness classification tasks [16, 6]. Instead of using the final LSTM hidden state as the sentence embedding, the Attention model performs a soft selection over the hidden states of the entire sequence to generate a sentence embedding. The attention mechanism is a simple affine layer activated with a tanh function before being passed to a softmax selection layer.

4. EXPERIMENTATION

Our dataset consists of follow-up utterances with ground truth annotations of device-directed or non-device-directed. The dataset contains roughly 240K utterances and is spilt into train, dev, and test partitions with 200K, 20K, and 20K utterances, respectively. The class split is roughly 5:1 in favor of the device-directed class in each partition, as non-device-directed speech is far less frequent.

4.1. Transfer Learning

While the dataset described above is not small, there is far more data available for first-turn interactions. We pre-train all models on a very large (2M) ASR false-wake dataset that contains first-turn device-directed speech and non-device-directed false-wake speech. False-wake speech are instances where the device mistakenly heard the wakeword. While first-turn and false-wake data is not exactly like follow-up mode data, the tasks are related enough that we consistently see improvement when pre-training on this data. We then fine tune the models on the proper follow-up dataset.

4.2. Model Training

The models were trained using SGD with the cross entropy loss function. We allowed for a decaying learning rate, and chose maximum and minimum learning rates using LR range tests [17]. For transfer learning, we pre-train with higher learning rates and fine-tune with smaller learning rates.

4.3. Model Evaluation

Table 5 shows the initial evaluation results for the candidate models, as well as an ablation test for our best model. We compare with the acoustic-only baseline model [5], but due to implementation and feature differences this model was trained once on a combination of our pre-trained and follow-up datasets. The DNN model is compared with and without the previous turn feature, as its average sentence embedding is unable to distinguish between turns. For the best model, we perform an ablation study to measure feature contributions and compare accuracy with and without transfer learning. Since our classes are imbalanced, we utilize the equal error rate (EER) metric.

The LSTM is able to outperform the average embedding model by a significant margin. The LSTM model with attention performs even better, though the improvement is much smaller. The best model with all features is able to reduce

Model	Features	EER %
ACOUSTIC-ONLY BASELINE [5] [†]	–	10.6
AVG-DNN	C,T	19.2
AVG-DNN	C,P,T	19.4
LSTM	C,P,T	9.2
LSTM+ATTN	C,P,T	**9.1**
LSTM+ATTN	C,P,T	9.1
	-C	13.7
	-P	9.5
	-T	14.8
LSTM+ATTN-NOTL	C,P,T	10.6

Table 5. Model evaluation with feature ablation for the best model. **C**: current utterance text, **P**: previous utterance text, **T**: token-level ASR confidences, [†]: trained in a different manner

Semi-supervised Pass	Test EER %	Dev Loss
0	10.4	0.176
5	10.1	0.175
10	10.0	0.173
15 [STOPPING POINT]	**9.9**	0.170
20	9.9	0.171

Table 6. Semi-supervised model evaluation

contribution to the overall score, notably in high-confusion posterior ranges.

5.2. Evaluation

Our unlabeled dataset U consists of 500K utterances. We select the LSTM without contextual features or attention; the unlabeled data does not contain previous turns and the attention model, likely due to hyperparameter issues, did not achieve good results during semi-supervised training. At each iteration we label the highest 1% of scores as DD and the lowest 0.2% of scores as NDD, to maintain the original class priors of the labeled dataset.

The stopping criteria is met once dev loss is no longer improving, in this case after 20 passes of semi-supervised training. We choose pass 15, the point of lowest dev loss, as our selected model. We evaluate this model, as well as several other versions, on the original test set and compare to the original model. Table 6 shows test set EER and dev set loss over select passes during semi-supervised training. Test EER improves 5% relative due to semi-supervised learning, and our stopping point based on the dev loss achieves the best EER.

6. CONCLUSIONS & FUTURE WORK

We have proposed a directedness classifier for spoken utterances based on acoustic and semantic lexical features. We have shown token-level semantic information, informed from word embeddings, can classify utterances with reasonable accuracy. Supplementing these lexical features with a lightweight acoustic feature and contextual information improves accuracy even further. Our results show that a recurrent model, such as an LSTM with attention, can utilize these features to reduce EER by 14% relative over an acoustic-only baseline model. Finally, we show that transfer learning and semi-supervised techniques can further improve the model without increasing the burden of data labeling.

Our future work will be along several dimensions. First, we plan to investigate more model architectures, such as CNN-LSTMs, FLSTMs, and Transformer-style models. For semi-supervised learning, we plan to experiment with much larger unlabeled datasets and more advanced graph-based techniques for exploration during training. We further plan to leverage contextual features and attention in our semi-supervised experimentation.

EER by 14% relative over the baseline model based on acoustic features.

The ablation study with the best model shows that all features add complementary predictive power to the model, albeit at different magnitudes. The token-level ASR confidences are the most important feature, quickly followed by the current turn text. The previous turn is less critical, likely due to the fact that only a subset of the dataset contains contextually-relevant utterances. Finally, our best model reduces EER by 14% relative over the same one trained without transfer learning *LSTM+Attn-NoTL*.

5. SEMI-SUPERVISED LEARNING

Labeling data is time consuming and costly. Semi-supervised learning approaches allow a model trained on labeled data to continue training on unlabeled data. Many approaches exist, but for simplicity we first investigate a self-teaching methodology. We use two models to aid in labeling like a co-training approach [18]. In this section we describe in detail our method and the results.

5.1. Method

We employ the self-teaching semi-supervised learning approach. In this method, first a model is trained in a standard supervised approach with labeled data L. Next, the trained model performs inference on a large set of unlabeled data U. The unlabeled cases where the model is highly confident U_L are labeled, removed from the unlabeled data and folded into training set. The model is then retrained with the new dataset $L + U_L$ and the process is repeated until stopping criteria is met.

We expand on the method by combining the scores of two distinct models, our proposed model and the acoustic-only baseline model, to reduce risk by labeling only those cases where the models have agreement. We apply a non-linear transformation to the acoustic model score to lower its

7. REFERENCES

[1] Ruhi Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 67–81, Jan 2017.

[2] Kyle Wiggers, "Baidu's dueros voice platform is now on 400 million devices," July 2019, [Online; posted 2-July-2019].

[3] Andrew Gebhart, "Google home hub listens longer with continued conversation," March 2019, [Online; posted 8-March-2019].

[4] Nick Statt, "Amazon adds follow-up mode for alexa to let you make back-to-back requests," March 2018, [Online; posted 9-March-2018].

[5] Sri Harish Reddy Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, and Björn Hoffmeister, "Device-directed utterance detection," *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. abs/1808.02504, 2018.

[6] Atta Norouzian, Bogdan Mazoure, Dermot Connolly, and Daniel Willett, "Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed," *CoRR*, vol. abs/1902.00570, 2019.

[7] Daniel S Reich, Felix Putze, Dominic Heger, Joris IJsselmuiden, Rainer Stiefelhagen, and Tanja Schultz, "A real-time speech command detector for a smart control room," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[8] Tim Paek, Eric Horvitz, and Eric K. Ringger, "Continuous listening for unconstrained spoken dialog," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2000.

[9] Heeyoung Lee, Andreas Stolcke, and Elizabeth Shriberg, "Using out-of-domain data for lexical addressee detection in human-human-computer dialog," in *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2013, pp. 221–229.

[10] John Dowding, Richard Alena, William J. Clancey, Maarten Sierhuis, and Jeffrey Graham, "Are you talking to me? dialogue systems supporting mixed teams of humans and robots," in *AAAI Fall Symposium: Aurally Informed Performance*, 2006.

[11] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tr, and Larry Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2012, pp. 334–337.

[12] Mark Gales and Steve Young, *Application of Hidden Markov Models in Speech Recognition*, Now Foundations and Trends, 2008.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[14] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin, "Advances in pre-training distributed word representations," in *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[15] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[17] Leslie N. Smith, "Cyclical learning rates for training neural networks," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2015.

[18] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, New York, NY, USA, 1998, COLT' 98, pp. 92–100, ACM.