BANGLA VOICE COMMAND RECOGNITION IN END-TO-END SYSTEM USING TOPIC MODELING BASED CONTEXTUAL RESCORING

Nafis Sadeq* *Shafayat Ahmed** *Sudipta Saha Shubha*†‡ *Md. Nahidul Islam**
Muhammad Abdullah Adnan*
{nafis, shafayat, sudipta, nahid.rimon}@ra.cse.buet.ac.bd, adnan@cse.buet.ac.bd
*Bangladesh University of Engineering and Technology (BUET)
†University of Virginia

ABSTRACT

In this work, we perform contextual rescoring using multi-label topic modeling to improve the performance of an End-to-End Bangla voice command recognition system. We use a hybrid of Connectionist Temporal Classification (CTC) and Attention mechanism in our End-to-End architecture. We use Recurrent Neural Network (RNN) as language model and Labeled LDA (Latent Dirichlet allocation) for contextual rescoring. Our experiments show that our rescoring method reduces Word Error Rate (WER) from 16.7% to 12.8% in Bangla voice command recognition task when the relevant context is provided. The system does not lose any performance when irrelevant context is provided.

1. INTRODUCTION

Voice command recognition task commonly involves an Automatic Speech Recognition (ASR) system with context-specific optimization. Context information for a specific smartphone user includes contact names, installed apps, songs, media files, location, recent search history, the content of the screen user is looking at, etc. These context information changes frequently so it is desired that the contextual model will be updated on-the-fly within the device.

Some notable work on contextual speech recognition include [1], [2], [3], [4], [5], etc. Google has incorporated contextual information with their state-of-the-art speech recognition system [6], [7], [8] and more recently with End-to-End speech recognition system [9]. All of the approaches used by Google are variations of n-gram based model for context detection. We propose a multi-label topic modeling approach for context detection which has several advantages over n-gram based approach. N-gram based approach is too rigid. It is not robust to synonymous, missing or misplaced words. All possible synonyms and n-gram variations need to be present in the contextual corpus. The topic modeling approach works on keywords which is more flexible and robust than the n-gram approach. A variable number of contexts can be easily handled with multi-label topic modeling.

Our contribution in this work is the following.

- We propose multi-label topic modeling based contextual rescoring for Bangla Voice Command recognition
- We consider a wide range of Bangla voice commands (for smart-phone, home appliance, automobile, etc.)
- Our rescoring system achieves WER of 12.8% when provided the context accurately. It outperforms all other existing voice command recognition systems in Bangla.

The rest of our paper is organized as follows. Our voice command domain is shown in section 2, system in section 3, contextual model in section 4, dataset in section 5, experiment results in 6 and Conclusion and future work in section 7.

2. PRELIMINARIES

Our primary objective was to cover all the voice assistant accessories that Bixby supports. Moreover, we consider the future scopes and select every possible domain in which we can recognize Bangla voice commands. We study the popular voice assistants - Google Assistant, Google Home, Amazon Alexa, Siri, Bixby, Cortana and more. We explore them and collect sentences from Smart-phone commands (System commands, Contacts, Media Player, Camera, Gallery, Messaging, Weather, Date, Alarm, Email, etc.), Home appliances (Smart TV, Fridges, Air-Conditioners, Computers etc.), Office work accessories (Projectors, Printers) and Automotive navigation applications (Vehicle routing, Utility Control) etc.

3. OUR SYSTEM

In this section, we describe our system in details.

3.1. System Overview

We use End-to-End ASR in our system. We use shallow fusion technique similar to the system described by [10] to incorporate the language model with the End-to-End architecture. Scores from CTC-Attention and language model are

[‡] work completed while being a Research Assistant at BUET

Thanks to Samsung R & D Institute Bangladesh for funding.

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

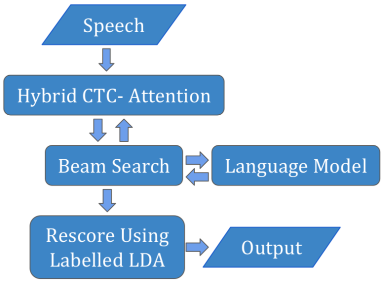

Fig. 1. System Overview

combined during beam search to generate a set of candidate hypotheses. Then we apply contextual rescoring on these candidates using Labeled LDA.

3.2. End-to-End Architecture

Our End-to-End architecture is based on the work of [11]. We use hybrid of CTC and attention mechanism. CTC and attention encoder networks share the same Bidirectional Long Term Memory Units (BLSTM). The encoder network had 4 layers with 320 BLSTM cells in each layer. The linear project layer has 320 cells. It is followed by each BLSTM layer. The decoder network has 1 layer. It has 320 unidirectional LSTM cells. For each audio frame, we use 40 MFCC features along with their first and second-order temporal derivatives. This gives us 120 features per frame. The shared encoder absorbs the input sequence into hidden states and the attention decoder generates the letter sequence.

During decoding with beam search, both attention scores and CTC scores are combined in the following manner. Let $p(o_n)$ be the probability of output label o_n at position n, given previous out labels and w_1 be the CTC weight.

$$\log p^{hyb}(o_n) = w_1 \log p^{ctc}(o_n) + (1 - w_1) \log p^{att}(o_n) \tag{1}$$

3.3. Language Model

The language model is trained on a large Bangla Text corpus. We experiment with both word-level and character-level Recurrent Neural Network (RNN). For word-level RNN, we use 1 hidden layer with 1000 LSTM cells. Most frequent 65000 Bangla words are considered in our vocabulary. For character level RNN, we use 2 hidden layers with 650 LSTM cells each.

3.4. Beam Search

We use shallow fusion technique to combine the language model scores into the End-to-End system [10]. Let, b be the width of beam search and v be the vocabulary size. At each step of beam search, b partial hypotheses are maintained by the system. In the next beam search step, each of these b hypotheses is extended by each of the tokens in the vocabulary. The total number of candidates becomes bv. For each of these candidates, the following score is calculated.

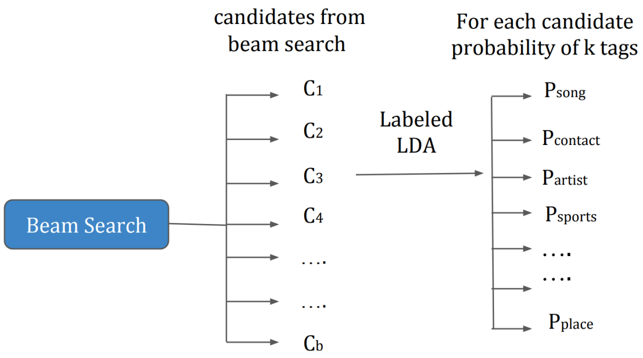

Fig. 2. Rescoring System

$$\log p(o_n) = \log p^{hyb}(o_n) + w_2 \log p^{lm}(o_n) \tag{2}$$

Here, $p(o_n)$ be the probability of output label o_n given the previous output labels, $p^{hyb}(o_n)$ be the score from hybrid CTC-attention system, $p^{lm}(o_n)$ be the language model score and w_2 be the language model weight. After calculating scores for each of the bv candidates, the top b candidates are considered for the next beam search step.

3.5. Contextual Rescoring

We use Labeled LDA for contextual relevance detection [12]. Regular LDA is an unsupervised algorithm that is not suitable for multi-label topic modeling. Unlike regular LDA, Labeled LDA allows the incorporation of a set of predefined topics. In our case, we consider each candidate sentence as a document and each contextual tag as a topic. We use label depth of 8 which is equal to the length of our contextual tags. We do not apply any dictionary pruning. Alpha and beta priors are set 0.1 and 0.01 respectively. Here, alpha represents document-topic density and beta represents topic-word density. We use a set of 37 contextual tags. We run training for 20 iterations.

After beam search, we have b candidates (assuming beam width of b) with their corresponding scores. First, we normalize the scores. Then, we apply topic modeling on each candidate. The output is a real-valued vector of length 37(i.e. the number of context tags). Each value represents the relevance of the sentence to that particular context. If the detected context matches with any of the on-device contexts, a bias is added to this candidate's score. Added bias is proportional to contextual relevance. We use a context weight w_3 to tune the context-sensitivity of the system.

Algorithm 1 Contextual Rescoring

```
$e2e \leftarrow$ Scores from End-to-End system
$rv \leftarrow$ Contextual relevance vector
$cl \leftarrow$ Current device contexts
$w_3 \leftarrow$ Weight of contextual bias
normalize(e2e)
for each c $\in$ candidateList do
    $rv = LabeledLDA(c)$
    for each r $\in$ rv do
        if $r > threshold$ and $r \in cl$ then
            $e2e[c] = e2e[c] + r \times w_3$
$output \leftarrow$ Candidate with maximum e2e score
```

4. CONTEXTUAL CORPUS GENERATION

We prepare a list of voice command templates from our domain study. These command templates contain entity tags. An example of a command template is '<contact> কে কল করো' (Call <contact>). Here, <contact> is an entity tag. We have a list of 1679 voice command templates containing around 20 entity tags. The entity tags include contact names, app names, number, time & date, song, artist, writer, book, place, movie, actor, food, gadget, team, player, company, etc.

Whenever the user adds a new contact, all the command templates containing the entity tag <contact> are populated with the new contact name and all the new sentences are added to the contextual text corpus. Similarly, when the user downloads a new song, all the command templates containing the entity tag <song> are populated and new sentences are added to the contextual text corpus. The contextual annotation of the newly formed sentences depends on the contextual annotation of the command template. Thus the contextual corpus gradually grows depending on the activity of the user. The labeled LDA system is periodically trained on the updated corpus. Personalized corpus of a particular user is fairly small, containing a few thousand sentences. It is possible to train the labeled LDA on a mid range smartphone within a minute.

5. DATASET

5.1. Text Corpus

The text corpus was prepared after extensive crawling from various popular Bangla websites. We crawl from around 42 websites and collect 10 million sentences. After collection of raw sentences, we use text cleaning to remove non-Bangla sentences, punctuation, alphanumeric characters, inconsistency, duplicates from the collected text. Later, we normalize these sentences. We convert numbers to text, handle abbreviations, manage special numeric expressions in Bangla, normalize decimal point & percentage symbol, consider contact numbers, date, etc.

5.2. Speech Corpus

We consider all publicly available Bangla speech corpus as well as prepared a speech corpus of our own. The largest publicly available speech corpus is provided by Google [13]. It contains 217902 utterances from 505 speakers. Among the speakers, 323 of them are men and 182 women. The size of the speech corpus is approximately 220 hours.

Bangla voice commands contain a set of technical words that are missing from all publicly available speech corpus. Also, the sentence structure of the voice commands is sometimes different from regular Bangla sentences. So we develop a speech corpus solely containing Bangla voice commands. We prepare an Android app for speech data collection following the approach by [14]. We are able to collect 28973 sentences from 56 speakers using this application. Among the speakers, 34 are men and 22 are women. The size of this corpus is around 50 hours.

6. EXPERIMENTS

6.1. Training Details

First, we train the hybrid CTC-attention based End-to-End system with publicly available Bangla speech corpus from Google [13]. Then we fine-tuned the system with our voice command corpus. The RNN based language model was trained with our Bangla text corpus containing 10 million sentences. The training of the End-to-End system takes around 72 hours and training of the RNN based language model takes around 18 hours. All experiments are done on a desktop with core i7 CPU, 16 GB RAM, Nvidia RTX 2070 GPU.

6.2. Test Set

Our test set contains 2000 utterances with 7 speakers. We manually annotated the context for these utterances to prepare a context annotated test set. We refer to it as positive test set. We also randomly annotate context of these utterances to prepare a negative test set. the purpose of the negative test set is to test whether the system's performance is affected with inaccurate context information.

6.3. Results

Table 1 shows Phoneme Error Rate (PER), Word Error Rate (WER) and Sentence Error Rate (SER) of our system in different setup. In our test set, the system using the character level RNN language model performs significantly better than the system using the word level RNN language model. The difference in performance is especially significant when test utterances contain out-of-vocabulary words. 700 out of 2000 test utterances contain one or more out-of-vocabulary words. For these utterances, WER was 26.6% and 46% for char-RNN

Language Model	Re-scoring	PER (%)	WER (%)	SER (%)
Word-RNN	None	6.9	27.9	44.9
	Trigram	6.4	25.7	42.4
	LLDA	6.0	23.8	40.3
Char-RNN	None	4.5	16.7	28.4
	Trigram	3.9	14.1	25.3
	LLDA	3.7	12.8	22.8

Table 1. Performance comparison

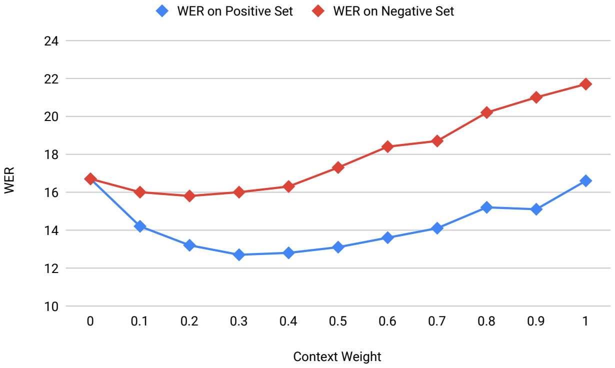

Fig. 3. Effect of Context Weight w_3

and word-RNN respectively. We tried larger word-RNN networks such as 2 layer,1000 LSTM cells and 2 layer,2000 LSTM cells. Overall WER were 27.2% and 26.9% respectively. Our rescoring method outperforms trigram based contextual rescoring method.

The performance of our system for different voice command categories can be found in table 2. In particular, regular system commands are recognized with very high accuracy (WER 7.6%) because they contain no out-of-vocabulary words. Highest WER (19.2%) is found in the case of random queries because they often contain out-of-vocabulary and out-of-domain input.

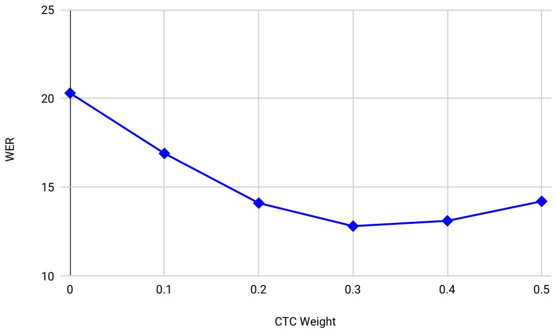

Fig. 4. Effect of CTC Weight w_1

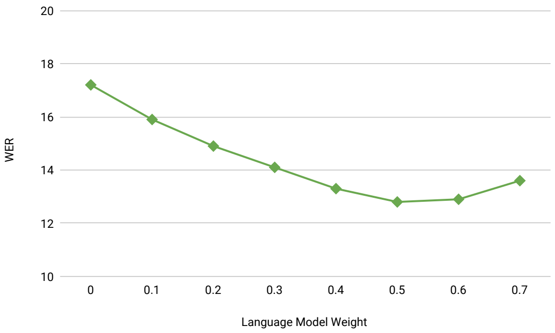

Fig. 5. Effect of Language Model Weight w_2

Category	WER
Regular System Commands	7.6
Numbers	9.2
Contacts	12.4
Place	13.5
Date and Time	15.7
Media	12.7
Random Queries	19.2

Table 2. WER for different Categories of Voice Command

Figure 3 shows the WER of the system for different values of context weight w_3 (algorithm 1). For the positive set, WER decreases upto $w_3 = 0.3$ then it starts to increase again. For the negative set, WER remains largely unaffected upto $w_3 = 0.3$ then it starts to increase almost linearly. For the experiment shown in Figure 3, we use CTC weight 0.3 and char-RNN language model with weight 0.5. Figure 4 and 5 shows the hyper-parameter tuning on validation set for CTC weight w_1 (equation 1) and language model weight w_2 (equation 2) respectively. We found best results using CTC weight 0.3 and language model weight 0.5.

7. CONCLUSION AND FUTURE WORKS

In this work, we use Labeled LDA for contextual rescoring to improve the performance of an End-to-End Bangla voice command recognition. Our experiments show that our rescoring method reduces WER from 16.7% to 12.8% in Bangla voice command recognition. In the future, we will try to enrich our contextual model to improve accuracy on out-of-vocabulary words.

8. REFERENCES

[1] Assaf Hurwitz Michaely, Mohammadreza Ghodsi, Zelin Wu, Justin Scheiner, and Petar Aleksic, "Unsupervised context learning for speech recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 447–453.

[2] Justin Scheiner, Ian Williams, and Petar Aleksic, "Voice search language model adaptation using contextual information," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 253–257.

[3] Tomas Mikolov and Geoffrey Zweig, "Context dependent recurrent neural network language model," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 234–239.

[4] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6186–6190.

[5] Suyoun Kim and Florian Metze, "Dialog-context aware end-to-end speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 434–440.

[6] Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno, "Bringing contextual information to google speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Leonid Velikovich, Ian Williams, Justin Scheiner, Petar Aleksic, Pedro Moreno, and Michael Riley, "Semantic lattice processing in contextual automatic speech recognition for google assistant," *Proc. Interspeech 2018*, pp. 2222–2226, 2018.

[8] Keith Hall, Eunjoon Cho, Cyril Allauzen, Francoise Beaufays, Noah Coccaro, Kaisuke Nakajima, Michael Riley, Brian Roark, David Rybach, and Linda Zhang, "Composition-based on-the-fly rescoring for salient n-gram biasing," `https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43816.pdf`, 2015.

[9] Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach, and Tara N Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search," in *Proc. of Interspeech*, 2018.

[10] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," in *Proc. of Interspeech*, 2017.

[11] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[12] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 248–256.

[13] Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha, "Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 52–55.

[14] Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro J Moreno, and Mike LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.