Design and Development of Spoken Dialogue System in Indic Languages

Shrikant Malviya
Department of Information Technology
Indian Institute of Information Technology Allahabad
Prayagraj, India
s.kant.malviya@gmail.com

Abstract

Based on the modular architecture of a task-oriented Spoken Dialogue System (SDS), the presented work focussed on constructing all the system components as statistical models with parameters learned directly from the data by resolving various language-specific and language-independent challenges. In order to understand the research questions that underlie the SLU and DST module in the perspective of Indic languages (Hindi), we collect a dialogue corpus: Hindi Dialogue Restaurant Search (HDRS) corpus and compare various state-of-the-art SLU and DST models on it. For the dialogue manager (DM), we investigate the deep-learning reinforcement learning (RL) methods, e.g. actor-critic algorithms with experience replay. Next, for the dialogue generation, we incorporated Recurrent Neural Network Language Generation (RNNLG) framework based models. For speech synthesisers as a last component in the dialogue pipeline, we not only train several TTS systems but also propose a quality assessment framework to evaluate them.

1 Introduction

Recently, substantial improvements in speech recognition performance have enticed the research community to build natural conversational interfaces in the form of a spoken dialogue system (SDS) (Jurafsky and Martin, 2019). This paper is concerned broadly with designing a complete spoken dialogue system in an Indic language scenario, i.e. Hindi. No significant work has been done earlier to promote the research and development of a Hindi spoken dialogue system. Hence, it becomes critical for the current work to address the issues and challenges unveiled for the Hindi language through introducing new datasets, methods and measures to build and evaluate all the integral modules of the Hindi SDS.

A typical SDS structure is based on a modular pipeline design connecting five principal components in a specific order (Pieraccini and Huerta, 2005): Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Manager (DM), Natural Language Dialogue Generation (NLDG) and Text-To-Speech Synthesiser (TTS). The work presented in this paper demonstrates how these components are developed individually and integrated at the end to develop a real-world spoken dialogue system in Hindi.

In a statistical spoken dialogue system, the aim is to replace each of the aforementioned components with a statistical model with parameters estimated from data (Young, 2002, 2010). The overall goal is to build a data-driven dialogue system with the ability to get improved over time and be perceived as behaving human-like by the users. The components of such systems are based on statistical methods, i.e. probabilistic distribution, neural network models, which allow them to handle uncertainty in both their inputs and outputs (Young et al., 2013; Zhang et al., 2001).

As the Hindi text contains lots of lexical/morphological ambiguities, therefore, it becomes a key challenge for DST and NLDG models to appropriately detect the DAs, understand the utterances and generate natural responses. Hindi is very rich in inflectional morphology. There is usually a limit of 8-9 inflected word forms of nouns in English (Yule, 2020), but in Hindi, it is more than 40 (Goyal and Lehal, 2008; Vikram, 2013). The way a language is spoken and written gets changed from place to place. It leads to the introduc-

Proceedings of the 18th International Conference on Natural Language Processing, pages 654–657
Silchar, India. December 16 - 19, 2021.

tion of variations where the meaning of a sentence is the same, but the way to express gets changed (Geeraerts et al., 2012).

Other language-related challenges that a Hindi SDS have to deal with are code-mixing (Ramanathan et al., 2009), hidden information (Miller et al., 1994), echo-words (Mohan, 2006), etc. Code-mixing is the mixing of more languages in the conversation. There are many cases in the corpus where the user had expressed some words from English during the conversation. (Example: "मुझे कम रेंज वाले रेस्तरां की तलाश है।" ("I am looking for low (cost) range restaurants.")). Here the word "रेंज" (range) is an English word that gives an indication of the cost.

2 Contributions

This research contributes at the following levels:

1. HDRS corpus: It raises the key research questions that underlie the *SLU* and *DST* module in building a Hindi dialogue system for the restaurant domain. Both traditional embeddings, i.e. *Word2Vec*, *GloVE* & *FastText* as well as *BERT* based embeddings are experimented.

2. A2CER: We incorporate the *advantage actor-critic with experience replay* (A2CER) algorithm (Wang et al., 2017) for dialogue policy learning which has recently been shown to be performing well on simple gaming environments and compare its performance with other state-of-the-art methods on a dialogue task.

3. Hindi NLG corpus: A corpus of unstructured input-output pair of *dialogue-act* (system's) and corresponding *natural response* is collected and released. The *RNNLG framework* based models are experimented on it.

4. Quality Assessment of TTS: A novel evaluation framework: *LBOE (Learning-Based Objective Evaluation)*, is developed for the quality assessment of various TTS systems. For the experiment, several "off-the-self" TTS systems: USS, HMM, CLU and DNN, have been trained from scratch.

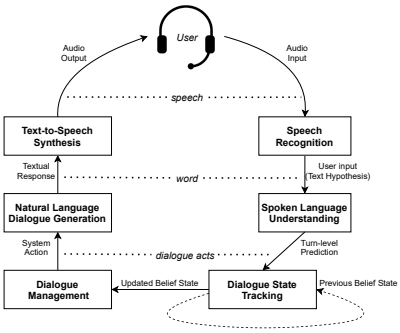

Figure 1: The pipeline of the core components in statistical spoken dialogue systems.

3 SILPA[1]: a Hindi SDS

We design our Hindi SDS by dividing it into five modules in a pipeline architecture (Pieraccini and Huerta, 2005) and connecting them in a specific order, as shown in Figure 1. The remainder of the section discusses the contributions specific to these modules.

3.1 HDRS: Language Understanding & State Tracking

For the empirical analysis of *language-specific* and *language-independent* challenges in dialogue state tracking, we release a dialogue corpus (HDRS) to train SLU/DST models in a *new language Hindi* with better annotations and high language-variability with significant corpus size (Malviya et al., 2021).

An SLU/DST component takes a sentence as input and maps it to an output dialogue act representing underlying semantics. For example, the utterance:

'मैं एक महंगा रेस्तरां खोज रहा हूँ जहाँ राजस्थानी खाना मिलता हो।'

can be represented as:

`inform(type=restaurant,price range=महंगा,food=राजस्थानी).`

3.2 Modelling Dialogue Management

We model the dialogue policy with RL approaches where the system's goal is to choose a sequence of system responses (actions) given the observed belief state achieving the maximum total reward, whereby the success of the dialogue mainly determines the reward. In this work, we have explored and investigated the current state-of-the-art methods

[1]SILPA (SILPAssistant): The name is based on our Lab's name SILP (Speech, Image & Langauge Processing) Lab

of policy optimisation for a task-oriented dialogue system, i.e. GP-SARSA, DQN, A2C. Inspired by (Wang et al., 2017), we present a new method that combines the strength of experience-replay in A2C (A2CER) policy learning for better dialogue modelling.

3.3 Natural Language Dialogue Generation

Obtaining the dialogue act from the dialogue manager, the Natural Language Dialogue Generation (NLDG) module transforms this abstract semantics notation (system dialogue act) back into a text representation (Singh et al., 2019). For example, the dialogue act:

`request(food)`

can be transformed to:

"आप किस प्रकार का भोजन खाना चाहेंगे? "

In our work, we have explored several state-of-the-art RNNLG-based models with discussing their performances on language-related (Hindi) challenges. All the models are experimented on our own Hindi dataset, collected on the restaurant domain.

3.4 Speech Synthesis & Quality Evaluation

At the last step in the SDS pipeline, the speech synthesis component converts the chosen text or the symbolic linguistic representation into a speech waveform. For the current study, we aim to cover leading TTS technologies as used in research as well as state-of-the-art commercial systems. Both TTS datasets, i.e. IIT-Madras, CMU, are used to build four types of unmodified "off-the-shelf" TTS systems: Unit selection synthesis (USS), Hidden Markov Model (HMM), Clustergen synthesis (CLU) and DNN synthesis (Tacotron 2). This forms the corpus and sets the background for our proposed 'LBOE' framework.

4 Dialogue Agent & Web Interface

We incorporated and adapted the multi-domain statistical dialogue System toolkit *PyDial-Toolkit* (Ultes et al., 2017) to build our dialogue agent "SILPAssistant". The **Agent** is the main component responsible for the dialogue interaction. The general architecture of the dialogue system with a speech interface is shown in Figure 2. The Agent can communicate to the user in both texts as well as

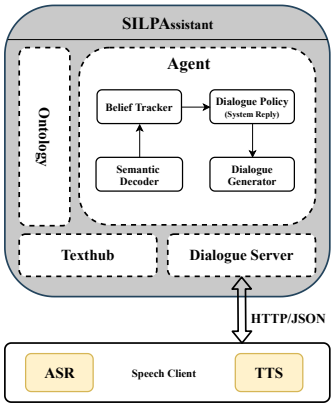

Figure 2: The general architecture of SILPA. The Agent resides at the core, and the interfaces Texthub, Dialogue Sever provide the link to the environment.

speech. For the text-based interaction, **Texthub** utility is provided, which simply connects the Agent to a terminal. To enable speech-based dialogue, the **Dialogue-Server** works as an interface between the Agent and the Speech-Client.

5 Conclusion & Future Studies

The current work has examined the challenges of developing a conversational system built upon native *Indian languages* for a real-world task. The original contributions of this thesis include: the development of an *HDRS* corpus on which various state-of-the-art SLU and DST models, i.e. *NBT*, *GLAD*, *GCE*, *GSAT*, *Simple-BERT* and *SUMBT*, are implemented and compared; the RNNLG models, i.e. *H-LSTM*, *SC-LSTM*, *MSC-LSTM* and *ENC-DEC*, have been experimented and used to train corpus-based NLDG module on a self-collected corpus in an Indic language Hindi; construction of dialogue policy with RL based approaches, i.e. GP-SARSA, DQN, A2C (Actor-Critic), A2CER (proposed), on the user-system act pairs generated by a user simulator; proposing a novel framework *LBOE* for quality assessment of a synthesised speech generated from various TTS engines, i.e. USS, HMM, CLU and DNN.

In the current work, we have explored a unimodal natural-language based dialogue scenario. As the human-to-human conversation is multimodal, involving various linguistic forms and non-verbal signals (Firdaus et al., 2021), a multimodal human-to-computer conversation should therefore be more intuitive.

References

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2021. Aspect-aware response generation for multimodal dialogue system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33.

Dirk Geeraerts, Stefan Grondelaers, and Peter Bakema. 2012. *The structure of lexical variation: Meaning, naming, and context*, volume 5. Walter de Gruyter.

Vishal Goyal and Gurpreet Singh Lehal. 2008. Hindi morphological analyzer and generator. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 1156–1159. IEEE.

Dan Jurafsky and James H Martin. 2019. Chatbots and dialogue systems. In *Speech and Language Processing (3rd draft ed.)*. Stanford University.

Shrikant Malviya, Rohit Mishra, Santosh Kumar Barnwal, and Uma Shankar Tiwary. 2021. HDRS: Hindi dialogue restaurant search corpus for dialogue state tracking in task-oriented environment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2517–2528.

Scott Miller, Robert Bobrow, Robert Ingria, and Richard Schwartz. 1994. Hidden understanding models of natural language. In *Proceedings of ACL*, pages 25–32.

Shailendra Mohan. 2006. Echo-word formation in hindi. *Indian Linguistics*, 67:119–126.

Roberto Pieraccini and Juan Huerta. 2005. Where do we go from here? research and commercial spoken dialog systems. In *Proceedings of SIGDIAL*, pages 1–10.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in english-hindi smt. In *Proceedings of ACL*, pages 800–808.

Sumit Singh, Shrikant Malviya, Rohit Mishra, Santosh Kumar Barnwal, and Uma Shanker Tiwary. 2019. Rnn based language generation models for a hindi dialogue system. In *International Conference on Intelligent Human Computer Interaction*, pages 124–137. Springer.

Stefan Ultes, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gašić, and Steve Young. 2017. PyDial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL*, pages 73–78. Association for Computational Linguistics.

Shweta Vikram. 2013. Morphology: Indian languages and european languages. *International Journal of Scientific and Research Publications*, 3(6):1–5.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. 2017. Sample efficient actor-critic with experience replay. In *Proceedings of ICLR*.

Steve Young. 2002. Talking to machines (statistically speaking). In *Seventh International Conference on Spoken Language Processing*.

Steve Young. 2010. Still talking to machines (cognitively speaking). In *INTERSPEECH*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

George Yule. 2020. *The study of language*. Cambridge University Press.

Bo Zhang, Qingsheng Cai, Jianfeng Mao, and Baining Guo. 2001. Planning and acting under uncertainty: A new model for spoken dialogue systems. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 572–579.