mTransDial: Multilingual Dataset for Transport Domain Dialog Systems (Poster)

Priyambada Ambastha
Indian Institute of Technology Hyderabad
cs18mtech11025@iith.ac.in

Maunendra Sankar Desarkar
Indian Institute of Technology Hyderabad
maunendra@cse.iith.ac.in

ABSTRACT

Task oriented virtual assistants or dialogue systems are being popular for different domains such as restaurant booking, weather update, flight booking etc. The efforts are supported by availability of large scale annotated conversational datasets for such domains. However, the same is not true for transport domain dialogue systems. Moreover, for such systems to be useful, they should be able to handle *natural queries* submitted by users. For countries like India where most of the people communicate in regional languages, it is important to have such systems support the regional languages. The existing datasets for transport domain are mostly monolingual in nature and support only English language. For countries like India, where people tend to speak multiple languages and have code-mixed conversations the existing systems and the datasets won't be of much use. To the best of our knowledge, there is no multilingual code-mixed dataset available for designing public transport related conversation systems. In this paper, we propose a code-mixed English-Hindi dataset to accelerate the development of transport domain conversational systems suitable for countries like India. Our dataset has multiple intents like: route finding, bus/train/cab finding, nearby place search, traffic alert queries, out of domain queries. We also provide initial baseline results for user intent identification using existing state of the art models on our dataset and a prototype to show the usability of the work.

CCS CONCEPTS

• **Computer systems organization** → **Natural Language Understanding**.

KEYWORDS

task oriented dialog systems, intent classification, code-mixed datasets, smart transportation

ACM Reference Format:
Priyambada Ambastha and Maunendra Sankar Desarkar. 2018. mTransDial: Multilingual Dataset for Transport Domain Dialog Systems (Poster). In *COMPASS .21: ACM SIGCAS Conference on Computing and Sustainable Societies, 28 June – 2 July 2021 (virtual).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Transportation is an important consideration in the design of sustainable smart cities. It has been observed and argued that problems like traffic congestion, air and noise pollution, and increasing fuel consumption have significant adverse impact on the quality of life in urban areas [1]. The increasing traffic volume also causes increase in the emission of green-house gases and contributes heavily towards the climate change [2]. Hence, efficient public transportation systems and mechanisms that can help people to make better decision making regarding transport systems [3] are desirable characteristics of any sustainable smart city.

Adoption of smart technologies that can help in transport-related decision making is hence an important concern of recent times. It has been argued that smart systems that can help people to take transport related decisions can be helpful in various aspects. For example, on-demand information regarding public transport schedule or current running status, information about congestion in different routes, accurate estimate of travel time along junctions, information about planned or unplanned events or incidents along traffic route etc. can be helpful for this purpose. Although these information can be obtained from multiple sources, such information is generally scattered over different special purpose applications and websites. People tend to use multiple mobile based applications or websites to get transit schedules for different modes of transport. Online schedules in many instances are clunky, involving multi-step input attributes to gain a simple answer. Many websites have pdf-based schedules or separate links to each metro/rail line. In some instances, there are separate sites for buses and trains. Hence, development of AI-powered conversational systems that can handle such kinds of information needs can be beneficial for implementation of urban transportation systems.

There are only a few such goal oriented road transport domain dialog systems [4] available till date. Moreover, they mostly consider schedule-related queries and hence are limited in their functionality. The system designed by [5] offers service for transport sector to *let bots manage the majority of common requests*, and is designed keeping the *benefit to the organization* in mind. There is a need to build transport related conversational systems that are user-centric and that cater to the different needs that the user may have even in multimodal (multiple transport types, general enquiries etc.) settings. We feel that lack of appropriate datasets to aid development of such systems is a limiting factor behind such development.

Developing goal oriented dialog systems require labeled training datasets for natural language understanding (NLU). NLU helps the system to know which actions have to be performed based on the user's query. There are multiple NLU datasets available for travel and flight booking domains, however, to our best knowledge there is no publicly available dataset in the public road transport domain.

COMPASS '21, 28 June – 2 July 2021, Virtual

ACM ISBN XXX-X-XXXX-XXXX-X/21/06…$15.00
https://doi.org/10.1145/1122445.1122456

Moreover, the publicly available datasets generally are monolingual and predominantly in English. To develop dialog systems for the countries where the native speakers use multiple languages and code-mixed utterances, having a code-mixed transport domain data is essential.

With the above motivation and to accelerate the development of road transport domain dialog systems we create a code-mixed English-Hindi dataset: *mTransDial* consisting of various kinds of road transport related queries. This dataset has been prepared using multiple seed datasets from different domains. Secondly, we provide initial baseline results on our dataset mTransDial using existing state of the art methods for text classification. Lastly, we built a prototype for road transport domain dialog system on a subset of our dataset using Rasa Framework.

2 THE PROPOSED DATASET

In this section we introduce a new dataset "mTransDial: Multilingual Dataset for Transport Domain Dialog Systems" for helping intent classification in transport domain dialogue systems. The dataset contains English-Hindi code-mixed queries for rail/road transport. It contains queries regarding nearby places, distance between places, route suggestions etc.

2.1 Seed datasets

Data annotation is a time consuming process. So, we decided to make use of the knowledge present in several existing datasets of relatable domain (here transport), and used them as seed datasets. We prepared a part of our dataset mTransDial using multiple existing datasets (ATIS [7], ChatbotCorpus [8], CLINIC150 [10]) from different domains. ATIS - Airline Travel Information System dataset is a standard benchmark dataset widely used for intent classification in the flight domain. We included queries from atis_flight (3666 queries) and atis_flight_time (54 queries) intents that talk about flight enquiry and flight time enquiry from source to destination. ChatbotCorpus contains 206 queries from a Telegram chatbot for public transport in Munich. It has two intents (Departure Time, Find Connection) out of which we have included FindConnection (128 queries) that have queries to find connection from source to destination in mTransDial. CLINIC150 contains 1200 out of scope queries i.e., queries that do not fall into any of the system's supported intents. It also has 150 intents from 10 domains that are in-scope. We have used 150 queries each from greeting, goodbye, thankyou, travel_alert, traffic, distance, directions intents in our dataset. To cater it to our purpose of code mixed English-Hindi queries we renamed each of the location entity mentions in the dataset to Indian locality names and mapped the intents from seed datasets to the desired intents in our dataset.

Countries like India where people tend to respond in multilingual utterances there is a strong need to develop a multilingual dialog system. If one wants to search for nearby restaurants can ask *"Can you please tell if mere aas pass koi acha restaurant hai"*, which means "Can you please tell if there is good restaurant nearby" in English . Thus, to make a multilingual dataset having code-mixed English-Hindi queries, we used Google Translate for translating the queries collected from the above data sources into Roman Hindi (transliterated Hindi). The translated version of the dataset is manually verified to eliminate the poorly translated queries from the dataset by the authors of the paper. This translated dataset contains a total of 6248 queries.

2.2 Dataset Characteristics

Figure 2 shows the distribution of number of examples in the translated dataset. We see a high class imbalance in this dataset as it has been collected from multiple seed datasets having different coverage over the intent classes. Figure 1 shows the word clouds generated from the dataset. Dataset contains text that are English to Hindi translations in transliterated form. Hence we see words like *tren* (originally "train" in English), *basen* (originally "buses" in English) etc. in the word cloud. There are many more words in the dataset that are translated like *samay* ("time" in English), *subah* ("morning" in English).

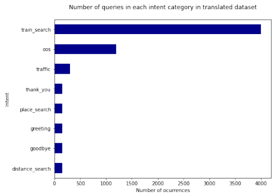


Figure 1: Query distribution in dataset

Figure 2: Wordcloud from Dataset

3 EXPERIMENTS

The following models were used for the task of intent classification on mTransDial:

- **Machine Learning & Deep learning models:** Support Vector Machine, Naive Bayes, Logistic Regression, Random forest models were used to classify intents from the translated dataset. We used TF-IDF feature vectors as input to our models. Convolutional Neural Networks and Long Short Term Memory were also used for evaluation.
- **Transformer based models:** BERT, Bidirectional Encoder Representations from Transformers have shown huge success on text classification tasks as it gives meaningful contextual representations of the input sentences. We can use this model for downstream intent classification task by finetuning it on the training set. In the BERT-BiLSTM method, we extracted embedding for each query in the dataset using BERT and passed it to a Bi-Directional LSTM. The BiLSTM representation was fed to 2 linear layers with relu activation function and dropouts followed by a softmax function to generate probabilities for each intent class. The next method BERT-ML transformer extracts BERT embeddings and passes it to machine learning classifiers (SVM, RandomForest & Logistic Regression).

The models are evaluated using 5-fold cross validation technique on mTransDial dataset. Accuracy and F1-Score were used as evaluation metrics. The results are presented in Table 2.

From Table 2 we observe that machine learning models like SVM and Logistic Regression with TF-IDF feature vectors performed

Table 1: Sample queries from the dataset. English translation of the queries are added for ease of understanding for the readers not familiar with the Hindi language

Intent	Query from the dataset	Query in English
Train_search	please mujhe kal mehadeepattanam se somajeeguda tak kee sabhee buses dikhaen	Please show me all the buses from mehadeepattanam to somajeeguda for tomorrow
Place_search	mujhe batao ki sabase kareebee coffee shop kahaan hai	Tell me where is the nearest coffee shop
Distance_search	begamapet jaane mein mujhe kitana time lagane vaala hai	how long is it going to take me to get to Begumpet
Traffic	kya husain saagar ke raaste mein traffic halka hoga	will traffic be light on the way to Hussain Sagar
OOS	bank kab tak khula hai	how long is the bank open until
Greeting	hellooo	hello
Thank you	dhanyawad	thank you
Goodbye	theek hai byee	okay bye

Table 2: Performance comparison of models on mTransDial

Model	F1-Score	Accuracy
TF-IDF+Multinomial NB	0.912	0.920
TF-IDF+Logistic Regression	0.957	0.966
TF-IDF+LinearSVC	0.971	0.970
TF-IDF+RandomForest	0.648	0.513
CNN	0.761	0.824
LSTM	0.763	0.826
Bert	0.971	0.972
MBert	0.972	0.974

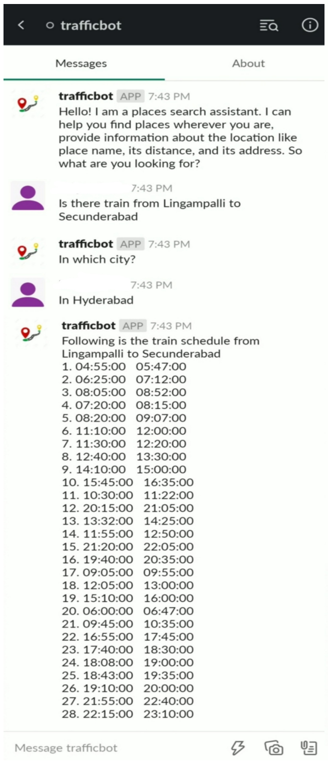
pretty well on the translated dataset. BERT and Multilingual-BERT gave similar results as our dataset contains Hindi text transliterated in English. Good performance by the ML models indicate frequent use of prominent terms in the data.

4 USABILITY IN ROAD TRANSPORT DOMAIN DIALOG SYSTEM

To assist in the development of road transport domain dialog system, we created a prototype that shows the usability of conversational systems in this domain. We use an open source Conversational system framework Rasa [9] to develop this prototype. The prototype is under development, and here we show a few examples involving English language queries.

We used a subset of our English data to build this prototype. Based on Rasa NLU, we annotated this data subset with the following entity information: Source_loc, Destination_loc, Near_loc, point_of_ interest. Source_loc and Destination_loc entities are to be extracted if the query is of Train_search or Distance_search intent class. For Place_search intent we extract Near_loc and point_of_interest entities to search for the nearby points of interest around the mentioned location. We used CRF entity extractor from Rasa NLU pipeline to extract the relevant entities.

If the classified intent is Place_search or Distance_search we call an external map based API (here, MapmyIndia[1]) to find the points of interest near to the Near_loc extracted or to find distance between Source_loc and Distance_loc. If the classified intent is Train_search we use GTFS (General Transit Feed Specification, a common format for public transportation schedules and associated geographic information) to fetch the relevant bus/metro/train schedule.



trafficbot APP 7:43 PM
Hello! I am a places search assistant. I can help you find places wherever you are, provide information about the location like place name, its distance, and its address. So what are you looking for?

7:43 PM
Is there train from Lingampalli to Secunderabad

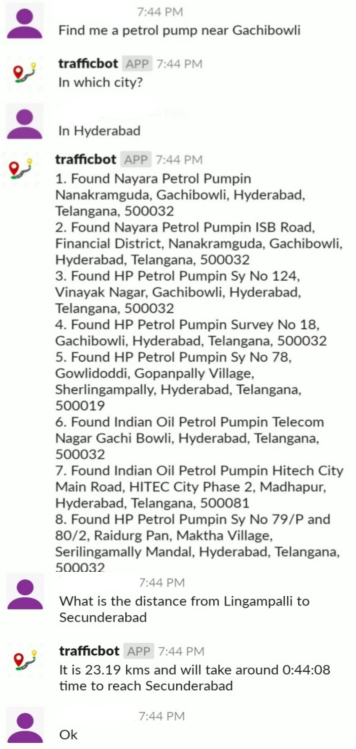
trafficbot APP 7:43 PM
In which city?

7:43 PM
In Hyderabad

trafficbot APP 7:43 PM
Following is the train schedule from Lingampalli to Secunderabad

```
1. 04:55:00   05:47:00
2. 06:25:00   07:12:00
3. 08:05:00   08:52:00
4. 07:20:00   08:15:00
5. 08:20:00   09:07:00
6. 11:10:00   12:00:00
7. 11:30:00   12:20:00
8. 12:40:00   13:30:00
9. 14:10:00   15:00:00
10. 15:45:00   16:35:00
11. 10:30:00   11:22:00
12. 20:15:00   21:05:00
13. 13:32:00   14:25:00
14. 11:55:00   12:50:00
15. 21:20:00   22:05:00
16. 19:40:00   20:35:00
17. 09:05:00   09:55:00
18. 12:05:00   13:00:00
19. 15:10:00   16:00:00
20. 06:00:00   06:47:00
21. 09:45:00   10:35:00
22. 16:55:00   17:45:00
23. 17:40:00   18:30:00
24. 18:08:00   19:00:00
25. 18:43:00   19:35:00
26. 19:10:00   20:00:00
27. 21:55:00   22:40:00
28. 22:15:00   23:10:00
```

Figure 3: Query output for Train_search intent



7:44 PM
Find me a petrol pump near Gachibowli

trafficbot APP 7:44 PM
In which city?

In Hyderabad

trafficbot APP 7:44 PM
1. Found Nayara Petrol Pumpin Nanakramguda, Gachibowli, Hyderabad, Telangana, 500032
2. Found Nayara Petrol Pumpin ISB Road, Financial District, Nanakramguda, Gachibowli, Hyderabad, Telangana, 500032
3. Found HP Petrol Pumpin Sy No 124, Vinayak Nagar, Gachibowli, Hyderabad, Telangana, 500032
4. Found HP Petrol Pumpin Survey No 18, Gachibowli, Hyderabad, Telangana, 500032
5. Found HP Petrol Pumpin Sy No 78, Gowlidoddi, Gopanpally Village, Sherlingampally, Hyderabad, Telangana, 500019
6. Found Indian Oil Petrol Pumpin Telecom Nagar Gachi Bowli, Hyderabad, Telangana, 500032
7. Found Indian Oil Petrol Pumpin Hitech City Main Road, HITEC City Phase 2, Madhapur, Hyderabad, Telangana, 500081
8. Found HP Petrol Pumpin Sy No 79/P and 80/2, Raidurg Pan, Maktha Village, Serilingamally Mandal, Hyderabad, Telangana, 500032

7:44 PM
What is the distance from Lingampalli to Secunderabad

trafficbot APP 7:44 PM
It is 23.19 kms and will take around 0:44:08 time to reach Secunderabad

7:44 PM
Ok

Figure 4: Response for Place_search & Distance_search

Figure 3 shows the output from our prototype for the query "nearby petrol pumps near Gachibowli". The system classified it into Place_search intent and extracted the relevant entitites i.e. *Near_loc: Gachibowli* and *point_of_interest: petrol pump*. Based on this, a call is made to map based API and the response is returned to the user. Figure 4 shows responses from the system for Train_search and Distance_search queries.

[1] https://www.mapmyindia.com

CONCLUSIONS

In this work, we present a multilingual code-mixed dataset containing traffic related user queries. The data can be used for query intent classification and can be helpful for developing transport domain conversational systems. We also present working examples from our transport domain dialog system that accepts user queries in English. An extension of this work will be to prepare multilingual code-mixed datasets that can be used to train other components of such dialog systems, and develop such a system that works end-to-end in multilingual manner.

ACKNOWLEDGEMENT

This work is done as part of the project “M2Smart: Smart Cities for Emerging Countries based on Sensing, Network and Big Data Analysis of Multimodal Regional Transport System”, supported by JST/JICA SATREPS, Japan (Grant Number: JPMJSA1606).

REFERENCES

[1] Hamurcu, Mustafa; Eren, Tamer. 2020. "Strategic Planning Based on Sustainability for Urban Transportation: An Application to Decision-Making" Sustainability 12, no. 9: 3589. https://doi.org/10.3390/su12093589 .

[2] Vajjarapu, Harsha, Ashish Verma, and Hemanthini Allirani. "Evaluating climate change adaptation policies for urban transportation in India." International journal of disaster risk reduction 47 (2020): 101528.

[3] Ambastha, Priyambada, and Maunendra Sankar Desarkar. "Incident Detection From Social Media Targeting Indian Traffic Scenario Using Transfer Learning." In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1-6. IEEE, 2020.

[4] M. Dharani, J. V. S. L. Jyostna, E. Sucharitha, R. Likitha and S. Manne, "Interactive Transport Enquiry with AI Chatbot," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1271-1276, doi: 10.1109/ICICCS48265.2020.9120905.

[5] Mindsay Public Transport Chatbot. https://www.mindsay.com/chatbot/public-transport. Accessed on: 9th April 2021.

[6] Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. "MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 5016-5026. 2018.

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

[8] Braun, Daniel, et al. "Evaluating natural language understanding services for conversational question answering systems." Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. 2017.

[9] Bocklisch, Tom, et al. "Rasa: Open source language understanding and dialogue management." arXiv preprint arXiv:1712.05181 (2017).

[10] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In Proceedings of EMNLP-IJCNLP.