Intent Detection and Slot Filling for Home Assistants: Dataset and Analysis for Bangla and Sylheti

Fardin Ahsan Sakib, A H M Rezaul Karim, Saadat Hasan Khan, Md Mushfiqur Rahman
Department of Computer Science, George Mason University *
{fsakib,akarim9,skhan225,mrahma45}@gmu.edu

Abstract

As voice assistants cement their place in our technologically advanced society, there remains a need to cater to the diverse linguistic landscape, including colloquial forms of low-resource languages. Our study introduces the first-ever comprehensive dataset for intent detection and slot filling in formal Bangla, colloquial Bangla, and Sylheti languages, totaling 984 samples across 10 unique intents. Our analysis reveals the robustness of large language models for tackling downstream tasks with inadequate data. The GPT-3.5 model achieves an impressive F1 score of 0.94 in intent detection and 0.51 in slot filling for colloquial Bangla. [1]

1 Introduction

Smart devices have become commonplace, establishing home assistants as indispensable fixtures in contemporary households. These voice-activated virtual companions adeptly manage an array of tasks, ranging from setting reminders to controlling room temperatures. The efficacy of home assistants in performing these tasks is closely intertwined with their underlying Natural Language Understanding (NLU) models, which enable seamless interactions in high-resource languages (Chen et al., 2019; Stoica et al., 2021; Antoun et al., 2020; Upadhyay et al., 2018). However, this advantage in NLU capabilities is not extended to low-resource languages (Stoica et al., 2019; Schuster et al., 2018), presenting a notable discrepancy. This discrepancy holds considerable significance, especially considering the global demand for home assistants and the extensive usage of low-resource languages, which have a substantial speaker base.

Bangla and Sylheti (Ethnologue, 2023), with 285 million native speakers combined, have rich cultural and colloquial nuances. Specialized datasets are needed to capture these intricacies as users prefer to interact with home assistants in their native languages, highlighting the research need (Bali et al., 2019).

The language understanding of home assistants is dependent on two key NLU tasks: intent detection and slot filling (Weld et al., 2022; Louvan and Magnini, 2020). Intent detection determines user actions, like playing music or checking the weather, while slot filling extracts specific details, such as song titles or locations. These tasks enable seamless human-device interactions, especially for home assistants.

Research on intent detection and slot filling primarily focuses on high-resource languages (Liu and Lane, 2016; Qin et al., 2021; Niu et al., 2019; Zhang et al., 2018). While there have been limited studies dedicated to the Bangla language (Bhattacharjee et al., 2021; Alam et al., 2021; Hossain et al., 2020), none of them have addressed the tasks of intent detection and slot filling in Bangla. Furthermore, these studies have not taken into account colloquial variants or closely related languages like Sylheti. This gap in research leaves a significant portion of the speaker base underserved.

This paper bridges this research gap with several notable contributions. Firstly, we introduce a comprehensive dataset encompassing 328 entries for intent detection and slot filling for each of the three languages – totaling 984 samples. These languages include formal Bangla, colloquial Bangla, and colloquial Sylheti. We further show a comparative study between generative LLMs and state-of-the-art language models for intent detection and slot filling.

2 Dataset

At the core of our exploration stands a meticulously curated dataset that is inspired by the SNIPS dataset (Coucke et al., 2018), which caters to the broad audience.

[1]The dataset and the analysis code can be found in the following directory: https://github.com/mushfiqur11/bangla-sylheti-snips.git

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 48–55

2.1 Dataset Size and Distribution

Originating from the 328 English samples present in the SNIPS dataset, our dataset underwent a manual correction phase to ensure that the English samples were of optimal quality. Then, we created three linguistically diverse variants, maintaining the same distribution across intent classes and slots as the original samples. These are:

1. **Formal Bangla:** This represents the standard version of the Bangla language, majorly used in contexts like official documents, news broadcasts, and literature. Formal Bangla tends to adhere strictly to grammatical rules.

2. **Colloquial Bangla:** An informal variant predominantly used in Bangladesh, colloquial Bangla resonates with everyday conversations of its people. While there are numerous dialects in different regions of Bangladesh, this form remains more or less consistent across the country. Colloquial Bangla is more flexible regarding syntax and incorporates a significant number of loanwords from English, Arabic, Persian, and other languages.

3. **Colloquial Sylheti:** A language with unique intricacies, Sylheti stands apart from Bangla and is spoken in the Sylhet region of Bangladesh and among diaspora communities. It's rich in expressions, proverbs, and idiomatic language that reflect the history and culture of the Sylhet region.

The curated dataset spans 10 distinctive intents. Each specific intent has a distinct set of slot categories. Figure 1 shows the number of samples for each intent and Figure 2 shows the fraction of slots that frequently occur for each intent, with respect to infrequently occurring slots.

2.2 Data Generation Process

The generation of our dataset was methodical and rigorous to ensure authenticity and accuracy.

Annotator Engagement

Four doctoral students were on board as annotators for our project. The initial phase involving the rectification of English data from the SNIPS dataset was a collaborative effort, with each annotator working on a distinct, non-overlapping segment. Subsequent phases involved two individuals fluent in Bangla for the Bangla datasets and two native Sylheti speakers for the colloquial Sylheti dataset.

Base Creation

The base dataset was created using the Bangla-T5 model (Bhattacharjee et al., 2023), a state-of-the-art English-to-Bangla translation tool, following the work of De bruyn et al.. The refined English samples served as the foundation to produce the initial Bangla translations for each sample. An auto-generated dataset comes with a myriad of issues. Therefore, these samples were manually re-translated and annotated with the auto-translations as the base.

Inter-Annotator Agreement

An essential step in ensuring the reliability of our dataset was to gauge the consistency between annotators. For each language variant, 28 randomly chosen samples were annotated independently by both designated annotators, followed by calculating their inter-annotator agreement (Table 1). This exercise helped us discern the degree of concordance and areas of divergence.

Consensus Building

Post the initial agreement calculation, a meeting was convened where the annotators discussed and reconciled their differences. This step was instrumental in ironing out inconsistencies and ensuring a unified approach going forward.

Blind Overlap

As the annotators progressed with data creation, a random 10% of the samples were earmarked for blind overlap. These served as a secondary check on inter-annotator agreement after dataset creation.

Independent Adjudication

After the final compilation of the dataset, each entry underwent a rigorous review by an independent adjudicator who had not previously worked on that particular language variant. This added an additional layer of scrutiny and quality assurance.

Inter-annotator agreement		
	Cohen's Kappa	Average BLEU
First 28 samples	0.42	0.43
Blind overlap (10%)	0.55	0.51

Table 1: There was an increase in annotator agreement before and after the annotator's meeting. This ensures the homogeneity of annotations in the dataset.

2.3 Ensuring Quality

Our data generation process, featuring multiple checks, blind overlaps, third-party reviews, and inter-annotator agreement stages, highlights our

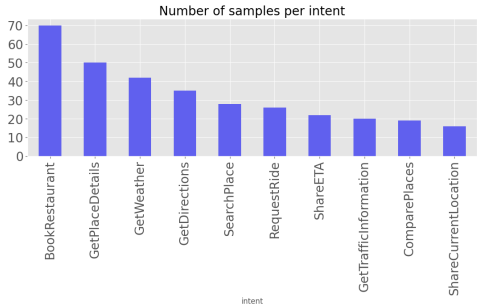

Figure 1: The number of samples for each intent varies, but they are fairly distributed, with 18 to 68 samples per intent.

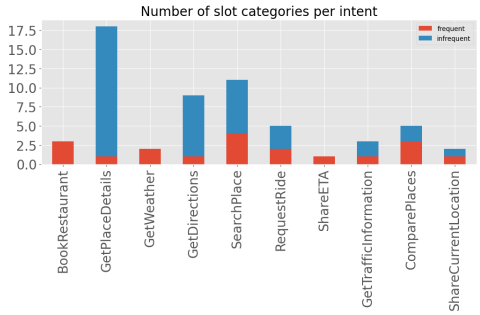

Figure 2: Slot categories appearing in at least 30% of the instances are marked as "frequent," while others are "infrequent." Despite varying slot categories per intent, frequent ones are evenly distributed.

commitment to quality. It minimizes biases and discrepancies that could result from a single annotator's viewpoint. The inclusion of an independent adjudicator in the final review further bolsters the dataset's integrity and reliability. Using a well-established dataset as the baseline ensures proper distribution of the data across different labels (Figure 1 and Figure 2).

3 Methodology and Experimental Setup

Our experiments were divided into four phases. In our initial experiment, we employed JointBERT (Chen et al., 2019), the state-of-the-art model in this domain, for both intent detection and slot-filling tasks. In our next experiment, JointBERT was retained for intent detection, while we explored the capabilities of GPT-3.5 (Generative Pre-trained Transformer) (Brown et al., 2020) model for slot filling. The third experiment fully utilized GPT-3.5 for both tasks. For our concluding experiment, we provided GPT-3.5 with the original intents and then analyzed its performance on the slot-filling task. The final experiment gives the raw result of slot-filling for the GPT model.

JointBERT leverages the BERT (Devlin et al., 2019) model to provide a unified approach encompassing both intent classification and slot filling by utilizing the representations from the pre-trained BERT model. We employed the default BERT tokenizer and maintained consistent parameters for all three languages. The utilization of these default settings and tokenization methods ensures an equitable and consistent evaluation across the languages.

GPT-3.5 (Generative Pre-trained Transformer) (Brown et al., 2020) model operates on the Transformer architecture and is adept at generating text resembling human language by predicting subsequent words or tokens in a sequence. GPT-3.5's deep contextual understanding is a result of extensive pre-training on a diverse corpus of textual data, encompassing various languages and linguistic intricacies enabling it to excel across a spectrum of NLP tasks (Goyal et al., 2022; Liu et al., 2021; Sakib et al., 2023; Kumar et al., 2020). We used GPT in a few-shot setting, passing 5 training samples along with the prompt. Rigorous prompt engineering was performed before settling on the two prompts for the two tasks. Figure 3 and Figure 4 show the final versions of the prompts used in the experimentations.

3.1 Experimental Setup

We divided each of the three datasets into training, development, and test sets using a standard 80-10-10 split. The JointBERT model was trained and evaluated on an A100 GPU, using a batch size of 8. We closely followed the setup provided by the original authors for this phase. For GPT, we used the OPENAI API with the "GPT-3.5-turbo" engine and set the token limit to 50.

4 Results

Tables 2 and 3 present the performance of the models we evaluated on our intent detection and slot-filling tasks. A clear pattern emerges: GPT-3.5 consistently outperforms JointBERT in both tasks.

While intent detection is generally more straightforward, JointBERT performs reasonably well in this aspect, although it doesn't quite match the exceptional performance achieved by GPT-3.5. However, when it comes to the more intricate task of slot-filling, JointBERT's performance falls significantly short, leaving ample room for improvement. In contrast, GPT-3.5 demonstrates its proficiency

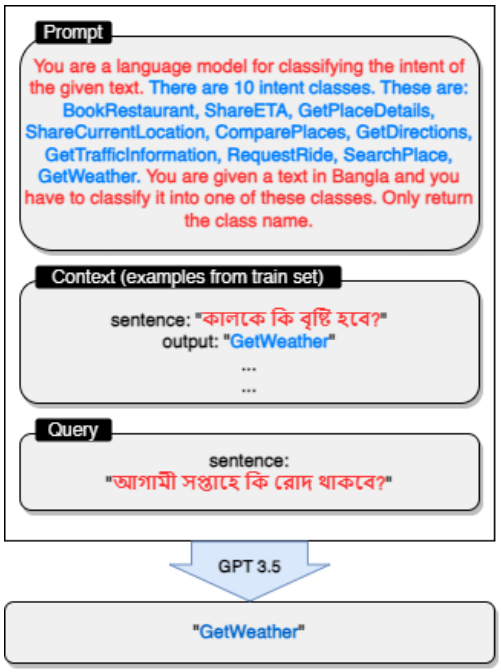

Figure 3: The figure illustrates how the input is formatted for the intent-detection task. A base-prompt is passed on to the GPT model. A few samples (5) from the training set are also passed as the context. From these sentence-output pairs, the LLM understands how the task needs to be solved. Finally, the current query is passed

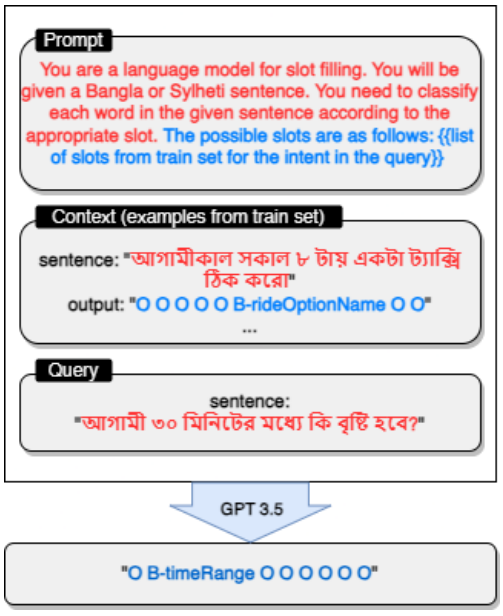

Figure 4: The input structure for the slot-filling task is quite similar to the intent detection task. The major difference is the prompt. For slot-filling, the set of possible slots is based on the intent type of the query. The intent type is obtained from a separate model and then from the train set, all possible slots for the given intent are fetched

Intent Detection (*Accuracy and F1 Score*)			
Models	Formal Bangla	Colloquial Bangla	Colloquial Sylheti
JointBERT	0.57 | 0.56	0.63 | 0.61	0.45 | 0.46
GPT-3.5	0.94 | 0.94	0.94 | 0.94	0.87 | 0.89

Table 2: While the performance of JointBERT is noteworthy for Bangla and its variants, the GPT-3.5 model excels across all metrics for all three datasets

Slot Filling (*F1 Score*)				
Slot Filling Model	Intent From	Formal Bangla	Colloquial Bangla	Colloquial Sylheti
JointBERT	JointBERT	0.14	0.11	0.07
GPT-3.5	JointBERT	0.43	0.45	0.52
GPT-3.5	GPT-3.5	0.45	0.51	0.57
GPT-3.5	Original	0.54	0.53	0.57

Table 3: The slot-filling task is separate from but dependent on the intent detection task. Intent needs to be passed to the model for good performance. In slot-filling tasks, GPT massively outperforms JointBERT

in handling the complexities of this task.

A significant reason behind GPT-3.5's superior performance is its broader exposure to diverse languages during training, including Bangla. JointBERT, conversely, hasn't been specifically trained on any Bangla dataset. This linguistic familiarity gives GPT-3.5 a clear advantage, enabling it to process and interpret Bangla's nuances far more effectively than JointBERT. The results underline the significance of using LLMs for low-resource languages, especially in scenarios where obtaining high volumes of training data for a particular downstream task is challenging.

5 Conclusion

In the era of smart devices, a home assistant's voice interfaces must resonate with the authentic linguistic intricacies of its users. Our research presents the first-ever dataset for intent detection and slot filling in Bangla and Sylheti, emphasizing their colloquial forms. This focus on colloquial forms bridges the often-overlooked gap between formal language models and the nuances of everyday speech. By championing colloquial forms, we ensure a voice interface that's more natural and attuned to genuine communication habits. Through rigorous data collection and validation, we have produced a high-quality benchmark dataset, providing a solid foundation for subsequent analyses and model evaluations. The comparative study between large lan-

guage models (LLM) like GPT-3.5 and non-LLMs underscores the remarkable capability of LLMs to excel even with minimal datasets, marking a considerable stride for underrepresented languages.

6 Limitations

While our research has made significant strides in understanding intent detection and slot filling for Bangla and Sylheti, like any study, it has its limitations. Our dataset, although carefully curated for the Bangla and Sylheti variants, is on the smaller side compared to established benchmarks. A precise and robust data generation process was prioritized, naturally limiting our data volume. We confined our evaluations to the JointBERT model and GPT-3.5. The pronounced difference in their performance deterred us from testing a broader range of models. Moreover, the dearth of optimized Bangla models for specific tasks posed challenges. An attempt with a Bangla BERT tokenizer didn't yield satisfactory outcomes, affecting the JointBERT's efficacy. As promising as our results are, they are tied to our specific dataset and context. Extending our findings to diverse settings or other languages requires further exploration, marking just the beginning of this exciting journey.

References

Masoud Akbari, Amir Hossein Karimi, Tayyebeh Saeedi, Zeinab Saeidi, Kiana Ghezelbash, Fatemeh Shamsezat, Mohammad Akbari, and Ali Mohades. 2023. A persian benchmark for joint intent detection and slot filling.

Firoj Alam, Arid Hasan, Tanvirul Alam, Akib Khan, Janntatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. *arXiv preprint arXiv:2107.03844*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. arxiv 2020. *arXiv preprint arXiv:2003.00104*.

Kalika Bali, Monojit Choudhury, Sunaya Sitaram, and Vivek Seshadri. 2019. Ellora: Enabling low resource languages with technology. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 160–163.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 714–723.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling.

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*.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent detection and slot filling for vietnamese. *arXiv preprint arXiv:2104.02021*.

Maxime De bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2022. Machine translation for multilingual intent detection and slots filling. In *Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22)*, pages 69–82, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ethnologue. 2023. Ethnologue 200: Languages of the world. Accessed on September 3, 2023.

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.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. Banfakenews: A dataset for detecting fake news in bangla.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833.

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

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *arXiv preprint arXiv:2011.00564*.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

Peiqing Niu, Zhongfu Chen, Meina Song, et al. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. *arXiv preprint arXiv:1907.00390*.

Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. *arXiv preprint arXiv:1909.02188*.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197. IEEE.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth annual conference of the international speech communication association*.

Fardin Ahsan Sakib, Saadat Hasan Khan, and AHM Karim. 2023. Extending the frontier of chatgpt: Code generation and debugging. *arXiv preprint arXiv:2307.08260*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Anda Stoica, Tibor Kadar, Camelia Lemnaru, Rodica Potolea, and Mihaela Dînşoreanu. 2019. The impact of data challenges on intent detection and slot filling for the home assistant scenario. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 41–47. IEEE.

Anda Stoica, Tibor Kadar, Camelia Lemnaru, Rodica Potolea, and Mihaela Dînşoreanu. 2021. Intent detection and slot filling with capsule net architectures for a romanian home assistant. *Sensors*, 21(4):1230.

Gokhan Tur, Li Deng, Dilek Hakkani-Tür, and Xiaodong He. 2012. Towards deeper understanding: Deep convex networks for semantic utterance classification. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5045–5048. IEEE.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6034–6038. IEEE.

Ngoc Thang Vu, Pankaj Gupta, Heike Adel, and Hinrich Schütze. 2016. Bi-directional recurrent neural network with ranking loss for spoken language understanding. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6060–6064. IEEE.

Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2020. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2018. Joint slot filling and intent detection via capsule neural networks. *arXiv preprint arXiv:1812.09471*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, volume 16, pages 2993–2999.

Zhichang Zhang, Zhenwen Zhang, Haoyuan Chen, and Zhiman Zhang. 2019. A joint learning framework with bert for spoken language understanding. *Ieee Access*, 7:168849–168858.

A Appendix

A.1 Related Work

Efforts to enhance datasets for intent detection and slot-filling within low-resource languages, such as Bangla and Sylheti in this context, commence with the intricate process of translating individual English lexemes extracted from established benchmarks like ATIS and SNIPS. Previous works in intent detection and slot filling for low resource languages(Dao et al., 2021; Akbari et al., 2023), have translated each English utterance to their respective languages. Recent works have shown that there are great performance achievements on intent detection and slot-filling tasks on datasets that have been derived from the SNIPS dataset (Weld et al., 2022; Qin et al., 2019; Wang et al., 2020), and this gives a reason to choose the SNIPS dataset over the ATIS dataset as it is a good starting point for a work with a language that has never been explored.

Spoken Language Understanding, a pivotal endeavor in the domain of task-oriented dialogue systems, encompasses the tasks of intent detection and slot-filling. Traditionally, these tasks were regarded as distinct domains in which significant progress was made (Tur et al., 2012; Ravuri and Stolcke, 2015; Mesnil et al., 2013; Vu et al., 2016). However, recent research has garnered notable attention by achieving remarkable advancements in performance through the concurrent learning of intent detection and slot-filling tasks (Zhang et al., 2018; Weld et al., 2022). In this section, we're primarily looking at how intent detection and slot-filling tasks are combined. We'll focus on two well-known strategies for this integration:

- A strategy devised through parameter sharing and the exchange of hidden states, utilizing a common BiLSTM/BERT encoder, along with two distinct decoders dedicated to intent detection and slot filling, on top of the shared encoder. (Chen et al., 2019; Xu and Sarikaya, 2013; Liu and Lane, 2016; Zhang and Wang, 2016).

- Another strategy, extending the initial approach to a more advanced level, involves the model acquiring an understanding of the relationships between slots and intent labels. This frontier has been explored in research in two distinct ways. Some studies (Goo et al., 2018; Li et al., 2018; Niu et al., 2019) have demonstrated the use of attention mechanisms to discern the correlation between the overarching intent context representation and the slot vectors generated by the encoder. Alternatively, other works (Qin et al., 2019; Zhang et al., 2019) have approached this by initially learning the representation of the utterance, which aligns with the representation of the global intent context, utilizing a self-attention mechanism. Subsequently, they join this representation with the encoder's vector outputs before feeding the combined vectors into the slot-filling decoder.

A.2 Examples from the dataset

Here we include a few examples from each of the datasets.

"আমার এয়ারবিএনবির কাছে একটি রেস্টুরেন্টে রাত ৮:৪৫ মিনিটের জন্য একটি টেবিল বুক করুন"- BookRestaurant- "O O O O O O O B-restaurant O O O O O O O O O O"

"আমি বাড়ি না আসা পর্যন্ত আমার বয়ফ্রেন্ডের সাথে আমার অবস্থান শেয়ার কর"- ShareCurrentLocation- "O O O O O B-contact I-contact I-contact O O O O O "

"আজ রাতে আমার ডিনারে যাওয়ার জন্য একটা উবার ডাকো" -RequestRide- "O O B-destination I-destination I-destination O O O B-rideOptionName I-rideOptionName O O"

Figure 5: Few examples from the Formal Bangla dataset. (Input sentence - Intent - Expected slots)

A.3 Prompts used for GPT

For the intent detection task we used the following prompt: *"You are a language model for classifying the intent of the given text. There are 10 intent classes. These are: BookRestaurant, ShareETA, GetPlaceDetails, ShareCurrentLocation, CompareePlaces, GetDirections, GetTrafficInformation, Re-*

"কফি ক্লাব কি সিপ কফি চেয়ে সস্তা?"- ComparePlaces- "B-place1 I-place1 O B-place2 I-place2 O O O"

"বদরুল কে আমার পৌঁছানোর সময় জানিয়ে একটা message পাঠাও"- ShareETA- "B-contact I-contact O O O O O O O O O "

"আমি আম্বারখানার যে রাস্তা দিয়ে আমার client meeting এ যাব সেইদিকে কি জ্যাম আছে?" -- GetTrafficInformation- "B-origin B-way I-way I-way I-way I-way O B-destination I-destination I-destination O O O O O O O O"

Figure 6: Few examples from the Colloquial Bangla dataset. (Input sentence - Intent - Expected slots)

"আমার মীরবাজার যাওয়ার লাগি ৫ মিনিট ওর বিত্রে এখটা ট্যাক্সি লাগব"- RequestRide- "O B-origin I-origin O O O O O O O O O B-rideOptionName O O"

"আমি বাড়িত জাইতাম কিলান?"- GetDirections- "O B-destination I-destination O O O O O O"

"আমি কিতা আমানুল্লাহর সামনে পার্ক খরতে ফারমু নি?" -- GetPlaceDetails- "O O O B-place I-place I-place O O O O O O O O O"

Figure 7: Few examples from the Sylheti dataset. (Input sentence - Intent - Expected slots)

questRide, SearchPlace, GetWeather. You are given a text in Bangla and you have to classify it into one of these classes. Only return the class name."

In this approach, we clearly outlined the potential intent classes, specified the input language as Bangla, and directed the model to solely return the class name. Such structuring was essential to elicit precise responses from the model.

For our slot-filling task, we utilized the following prompt: *"You are a language model for slot filling. You will be given a Bangla sentence. You need to classify each word in the given sentence according to the appropriate slot. The possible slots are as follows: list of possible slots extracted from the train set (based on the training intent)"*

We equipped the model with both the potential slots and their associated intent. Notably, the performance fluctuated depending on the source of the intent— GPT-3.5, JointBERT, or the Original dataset.

A.4 Inter-annotator metrics

In order to assess inter-annotator agreement, this study utilized two primary evaluation metrics: Cohen's Kappa and Average BLEU.

Cohen's Kappa provides a statistical measure of agreement between two annotators, while accounting for the possibility of chance agreement. Specifically, it involves calculating the actual observed agreement between the annotators and comparing that to the level of agreement that would be expected by random chance. Cohen's Kappa expresses the ratio between these two values as a score ranging from 0 to 1, with higher scores indicating greater reliability.

Average BLEU (Bilingual Evaluation Understudy) is a commonly employed metric for evaluating machine translation outputs by comparing them against one or more reference translations. It analyzes the co-occurrence of n-grams between the translated text and human reference texts to produce a score reflecting the quality and fluency of the translation. Taking the average BLEU score across multiple translations provides an overall indicator of the fidelity of the translations with respect to the reference materials.

Together, these two metrics enable analysis of both the reliability of individual annotators via Cohen's Kappa and the accuracy and fluency of translations via Average BLEU in relation to trusted references. The combination provides a robust means of evaluating key aspects of annotation quality for this study.