

THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF COMPUTER SCIENCE

GUIDE: A Dataset and Study of Context for Intent Recognition in Asymmetric Guiding
Dialogues

Jincheng Lyu
FALL 2025

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Computer Science
with honors in Computer Science

Reviewed and approved* by the following:

Rebecca Jane Passonneau
Professor of Computer Science & Engineering
Thesis Supervisor & Honors Adviser

Sencun Zhu
Professor of Computer Science and Engineering
Thesis Faculty Reader

Abstract

Intent recognition in situated, goal-directed dialogue remains challenging because utterances are highly context dependent and speakers often possess asymmetric knowledge. This thesis studies intent classification in instructional interactions by introducing GUIDE (Growing Understanding through Interactive Daily Experience), a curated dataset of 447 transcripts collected from VirtualHome-based simulations of 10 household activities, each instantiated in four controlled variants to elicit guidance, confusion, and error correction. GUIDE includes fine-grained, role-sensitive intent labels (e.g., Inform, Request Information, Request Action, Casual, Greetings), distinguishing human guidance from the alien learner’s knowledge reports and confirmations.

We first establish that label quality is pivotal: manual standardization of intents across 23,195 utterances improves BERT-base accuracy from 0.64 (uncleaned) to 0.70 (cleaned) under identical settings. We then evaluate pre-trained models (ALBERT, T5, BERT-base, BERT-large) and analyze the effect of context size and input construction. Short multi-turn context (target + 1 prior turn) consistently helps, raising accuracy to 0.74; end padding stabilizes performance (0.75); compact role encoding and input pruning further reduce token overhead; and summing token embeddings across utterances yields a best fine-grained accuracy of 0.76. Aggregating labels into high-level categories substantially simplifies the task, achieving up to 0.87 with BERT-large (1 epoch).

Our findings show that (i) careful curation, (ii) short, local context, and (iii) lightweight input engineering materially benefit intent recognition in instructional dialogues. The resulting single-stage, supervised classifier operates offline and does not perform dialogue management or action generation. We conclude with limitations (domain scope, transcript variability) and outline directions for multimodal integration, hierarchical/online classification, and broader data collection.

Table of Contents

List of Figures	iv
List of Tables	v
Acknowledgements	vi
1 Introduction	1
1.1 Background and Problem Statement	1
1.2 Significance of the Problem	2
1.3 Proposed Approach: Dataset and Classifier	2
1.3.1 Dataset Creation	2
1.3.2 BERT Classifier for Intent Recognition	3
1.4 Thesis Objectives and Structure	3
2 Related Work	4
2.1 Summary of Related Datasets	4
2.2 MultiWOZ Dataset	5
2.3 Deal or No Deal	6
2.4 Decoupling Strategy and Generation in Negotiation Dialogues	6
2.5 MindCraft	7
2.6 Analysis of Dialogue in Human-Human Collaboration in Minecraft	7
2.7 SIMMC 2.0	8
2.8 TEACH	8
2.9 Emotional Support Dialogues	8
3 Dataset Description	9
3.1 Overview of the Dataset	9
3.1.1 Intent Labels	10
3.2 Dataset Acquisition and Preprocessing	15
3.2.1 Additional Preprocessing Considerations	15
3.2.2 Challenges and Limitations of the Dataset	16
4 Experiments on Context Size	17
4.1 General Performance and Model Selection	17
4.1.1 Impact of Data Cleaning	18

4.1.2	Comparison of Pre-trained Models	18
4.1.3	Hyperparameter Tuning: Training Epochs	19
4.1.4	Performance on High-Level Intent Categories	20
4.2	Exploration of Multi-Utterance Context	22
4.2.1	Results with Two Utterances	22
4.2.2	Results with 3 Utterances	22
4.3	Exploration of Padding Methods	22
4.4	Further optimization	23
4.4.1	Optimization of Training hyperparameters	23
4.4.2	Exploration of Token Length and Role Encoding	24
4.4.3	Token Summing for Multi-Utterance Representation	24
4.5	Summary	25
5	Conclusion	26
5.1	Summary of Contributions	26
5.2	Key Findings	26
5.3	Discussion	27
5.4	Limitations and Threats to Validity	28
5.5	Future Work	28
5.6	Concluding Remarks	29
	Bibliography	30

List of Figures

3.1	10 scenes in GUIDE Dataset	11
-----	--------------------------------------	----

List of Tables

2.1	Eleven human-machine dialogue datasets illustrating the variety in domains, modalities, and number of intent types	5
4.1	Example of misunderstood label	18
4.2	Accuracy Comparison across models, and two settings for training epochs	19
4.3	Per-Intent Metrics for High-Level Categories (BERT-large and T5)	21
4.4	Summary of Accuracy Across Key Configurations	25

Acknowledgements

Chapter 1

Introduction

1.1 Background and Problem Statement

Traditional NLP approaches, such as rule-based systems or early machine learning models, rely heavily on syntactic and semantic parsing but struggle with the subtleties of pragmatics. Deep learning models, including pre-trained language models (PLMs) like BERT (Bidirectional Encoder Representations from Transformers) [1], or large language models (LLMs) like those in the GPT family, perform extremely well for many tasks that require language understanding or language interaction, but cannot handle situated, interactive language understanding and generation. Applications such as virtual assistants (e.g., Siri or Alexa), chatbots, and automated customer service systems frequently misinterpret user queries, leading to frustration, inefficiency, and reduced trust in AI technologies. For example, a user saying “turn it on” might intend to activate a device, start a conversation, or even metaphorically encourage someone, but without robust intent classification, AI responses can be irrelevant or erroneous. There are some previous approaches to work on conversation; one approach is virtual world, which creates entire new rules of action, such as in the Diplomacy game where representations for utterances from deep learning are combined with reinforcement learning [2], while the other simulates the real world, such as in the Seagull environment that uses traditional NLP methods [3].

A critical feature of systems for situated dialogue like SeaGull or Cicero is the requirement that they recognize the underlying intent of an utterance, which is not what PLMs or LLMs are designed to do. Intent recognition is challenging because the same intent can be expressed in many different ways, and the same utterance can represent different intents, depending on the context. Therefore our investigation of classification performance also includes the question of how much of the context to use. To conclude, the core problem this thesis addresses is: **How well can a classifier based on a PLM perform in classification of intents in conversational phrases that vary across different contexts?** Without a mechanism to account for these variations, AI remains

limited in its ability to engage in natural, human-like interactions. This gap highlights the need for specialized datasets and classifiers that explicitly incorporate contextual diversity to improve intent recognition accuracy.

1.2 Significance of the Problem

Intent recognition is not merely an academic pursuit; it has profound implications for real-world applications. In healthcare, for instance, chatbots must distinguish between a patient’s casual complaint and a urgent distress signal to provide appropriate responses. In education, AI tutors need to detect whether a student’s question stems from confusion, curiosity, or disengagement. Similarly, in customer service, misinterpreting intent can lead to escalated issues or lost business. Intent misclassification remains a significant reason for AI failures, contributing to substantial economic losses due to poor customer experiences.

Current datasets for intent recognition, such as ATIS (Airline Travel Information System) [4], are often domain-specific and lack the breadth of contextual variations found in everyday conversations. ATIS falls into the category of customer service interactions through an interface to a database. Such systems typically focus on scripted queries rather than dynamic, ambiguous phrases. Moreover, while many are gathered from real-world situations, these are highly restricted (e.g., database queries), enabling the collection of massive amounts of data but limiting their scope to non-physical contexts without elements like navigation or object manipulation. This thesis bridges utilizes a novel dataset and classification approach, contributing to the broader NLP community by advancing context-aware AI systems.

1.3 Proposed Approach: Dataset and Classifier

To investigate the problem of contextual intent recognition, this thesis introduces a supervised learning framework centered on a custom dataset and a BERT-based classifier. The key insight is that training AI on diverse, contextually rich examples can help us understand contextual factors that influence intent recognition, much like how humans infer meaning from situational cues.

1.3.1 Dataset Creation

A critical component of our investigation is the development of a specialized dataset designed to capture intent ambiguities in goal-directed, situated interaction. We created a series of controlled scenarios simulating a human guiding an “alien” (a fictional entity unfamiliar with human norms) how to perform everyday activities. These activities include using a dishwasher, reading a book, watching television, and similar tasks that mimic real-world instructional dialogues. The scenarios were generated as videos in the VirtualHome environment, which does not include any form of language interaction (text or audio), thereby focusing the subsequent transcript generation on inferred textual dialogues.

For each video setup (collected through collaborations with Pennsylvania State University NLP researchers; more detail in Chapter 3), we collected multiple transcripts from human participants. These participants worked in pairs, with each pair role-playing the human and alien separately to

add verbal interactions to the physical simulation (e.g., what the human might say to the alien or how the alien might respond). This resulted in a dataset comprising varied multiple ways to verbalize the inferred intents of the human and alien in a given simulated interaction of the same visual scene, emphasizing differences in intent between human and alien perspectives. For example:

- In a “using a dishwasher” scenario, a human transcript might focus on prompting the alien to perform a task that needs to be completed (“Actually, wait, while we’re here, can you grab that plate? We’ll need to wash it anyway.”), while an alien perspective could convey confirmation or corrective intent (“Yeah, that should be fine. Now we can start cooking.”).
- Similarly, in a “reading a book” scenario, a human transcript might be motivational (“Keep turning the pages; you’ll get the hang of it”), while an alien one informs the human of their current condition (“I have finished reading the second book now too.”).

This approach generates patterns where identical or similar phrases appear in different contexts, highlighting intent shifts (e.g., the phrase “turn it on” could mean activating the TV or encouraging the alien to engage). The dataset consists of 447 transcripts, with labels for intent categories such as inform information, request information etc. By including multiple transcripts per setup, the dataset captures variations in how annotators attribute intents to the same scenario and express those intents through different phrasings, thereby modeling the variable relationship between wording and associated intents. This complements the contextual variability already provided by the diverse scenarios, addressing the limitations of existing benchmarks.

1.3.2 BERT Classifier for Intent Recognition

The dataset serves as the foundation for training a BERT classifier in a supervised learning paradigm. BERT, a pre-trained transformer model, excels at capturing bidirectional context in text, making it a useful baseline candidate for intent classification [5]. We fine-tune BERT on our labeled transcripts. The model learns to predict intent labels, enabling it to generalize to unseen conversational phrases.

This method outperforms baseline approaches by incorporating multimodal-inspired data (visual scenarios informing text), thus improving accuracy in ambiguous cases. Preliminary experiments (detailed in Chapter 4) show promising results, with the classifier achieving highest of 0.76 overall accuracy

1.4 Thesis Objectives and Structure

The primary objective of this thesis is to demonstrate that a contextually diverse dataset, combined with a BERT classifier, can effectively enable a classifier to recognize intents in goal-directed dialogues in a simulated physical environment. Secondary objectives include evaluating the dataset’s quality, comparing our approach to state-of-the-art models, and discussing implications for future NLP research.

Through this work, we aim to contribute a practical tool for advancing context-aware AI, paving the way for more intuitive human-machine interactions.

Chapter 2

Related Work

In this chapter, we review key datasets and studies in task-oriented dialogues, negotiation, collaboration, and teaching scenarios. These works enable a comparison with our research on intent classification in human-alien guidance dialogues. We highlight similarities and differences, emphasizing how our approach prioritizes granular intent processing—such as handling contextual ambiguities in instructional utterances—over broader dialogue state tracking.

To frame these comparisons, we briefly introduce our GUIDE dataset (Growing Understanding through Interactive Daily Experience). GUIDE focuses on asymmetric task completion interactions, where a human provides guidance to an “alien” (a novice entity) on everyday tasks carried out by simulated agents in a simulated world. It features transcripts of utterances made by the human-alien pairs annotated with intent labels across categories like Inform, Request Information, and Request Action, highlighting a gap in existing datasets that often involve symmetric collaboration or transactions [6]. There are few if any situated dialogue datasets where the overarching dialogue goal is to provide verbal assistance on ordinary tasks.

Task-oriented dialogues involve goal-directed interactions, such as service bookings, where systems track user intents for efficient fulfillment [4]. Negotiation scenarios emphasize bargaining, like resource division, focusing on conflict resolution [7]. Collaboration datasets capture joint problem-solving in shared environments, prioritizing coordination [8]. Teaching scenarios, most akin to GUIDE, center on knowledge transfer from expert to novice but typically emphasize AI task execution rather than deep intent analysis in utterances [6].

2.1 Summary of Related Datasets

Table 2.1 summarizes key datasets, highlighting modalities, sizes, goals, domains, and dialogue acts.

Dataset (Domain)	Modality	N Dial.	Goal Type	Intent Types
MultiWOZ[9] (restaurants . . . ; N=7)	text	10.0K	Customer Service	13
Deal or No Deal[7] ({books, hats, balls})	text	5.8K	Negotiate	NA
CraigslistBargain[10] (6 categories)	text	6.6K	Buy-Sell	Implicit
MindCraft[11] (Minecraft mining)	text, images	2.1K	Asymmetric knowledge collaboration	NA
Coll. Garden Task[12] (Minecraft garden)	text, images	0.5K	Collaboration	NA
SIMMC 2.0[13] (Furniture, fashion)	text, images	11.2K	Shopping Asst	4 + 5
TEACH-DA[6] (12 household activities)	text, video simulations	4.3K	Leader/follower	18
ESConv[14] (5 problem types)	text	1.0K	Emotional support	8
EmpatheticDialogues[15] (User select problems)	text	24.8K	Emotional support	NA
MDC[8] (Build Minecraft structures)	text + minecraft world	0.6K	Collaborative Game	6
GUIDE	text	18.4K	Intent Classification	39

Table 2.1: Eleven human-machine dialogue datasets illustrating the variety in domains, modalities, and number of intent types

2.2 MultiWOZ Dataset

The MultiWOZ dataset [9] is a large-scale collection of goal-directed, multi-domain dialogues for task-oriented systems. It comprises over 10,000 human-human written conversations simulating tourist-clerk interactions across seven domains, such as hotels, restaurants, and attractions. Annotations include dialogue acts and belief states, primarily aimed at evaluating dialogue state tracking in customer transactions like booking reservations. Subsequent versions, such as MultiWOZ 2.1 to 2.4, refined annotations to improve consistency and address initial data quality issues.

Task-oriented dialogue systems, as exemplified by MultiWOZ, are typically mediated by a computer interface rather than direct agent-to-agent communication. They are commonly applied to customer-service or database-driven interactions, where the agent and human engage solely through an interface (e.g., text or voice queries) without reference to a real or simulated physical world. Tasks in these systems are narrowly defined, such as booking a hotel or restaurant reservation, and multi-domain aspects cover variations across task types (e.g., combining hotel and taxi bookings in a single dialogue). MultiWOZ addresses key challenges in this area, including the need for scalable, multi-domain benchmarks to train models that handle complex user queries, track evolving dialogue states, and generate appropriate responses in transactional settings. Evaluation

in MultiWOZ focuses on metrics like joint goal accuracy and slot-filling precision, emphasizing the system’s ability to maintain accurate belief states across turns.

While MultiWOZ shares goal-directed elements with our GUIDE dataset — such as structured interactions aimed at achieving specific outcomes — the underlying goals differ significantly, influencing intent types and evaluation approaches. MultiWOZ prioritizes transactional intents (e.g., querying availability or confirming bookings) in symmetric, service-oriented dialogues, evaluated through state-tracking metrics that assume a database backend. In contrast, our approach centers on asymmetric guiding dynamics in human-alien scenarios, where intents involve instructional guidance, confusion resolution, and contextual adaptations (e.g., handling “lost moments” or errors in a simulated physical world). This leads to distinct intent categories, such as motivational or corrective labels, and evaluation focused on context-aware intent recognition rather than transactional state tracking, better suited to real-world knowledge transfer applications.

2.3 Deal or No Deal

The Deal or No Deal dataset from Lewis et al. (2017) [7] involves end-to-end learning for negotiation dialogues. It contains 5,000 crowdsourced dialogues based on 2,000 scenarios with items like books, hats, and balls, each with assigned values. Two humans negotiate to divide items, focusing on agreement or disagreement. This is similar to human-alien discussions in our setup, but it emphasizes debating for deal agreements, unlike our focus on guiding and intent classification.

Unlike many of the other works cited in this chapter, Deal or No Deal employs a fully end-to-end model, training neural networks to handle both linguistic generation and strategic reasoning directly from raw dialogue inputs to outputs, without relying on annotated intermediate states or modular components. For contrast, datasets like MultiWOZ [9] and TEACH [6] support modular task-oriented systems where components such as intent classification, state tracking, and response generation are trained separately, often requiring explicit annotations for dialogue acts or belief states to evaluate sub-tasks independently. Similarly, CraigslistBargain [10] decouples strategy from generation, treating them as distinct modules rather than a unified trainable process.

Like Deal or No Deal, which integrates negotiation strategy and language output in a complete loop for role-asymmetric (buyer vs. seller) bargaining tasks, our GUIDE-based classifier is tailored to asymmetric interactions, but serves as a modular component for recognizing context-sensitive guidance intents (e.g., corrective or motivational) in teacher–learner dynamics within situated household scenarios.

2.4 Decoupling Strategy and Generation in Negotiation Dialogues

He et al. (2018) present the CraigslistBargain dataset for decoupling strategy and generation in negotiation dialogues. It includes 6,600 crowdsourced dialogues where participants act as buyers or sellers of real-world products from Craigslist categories like housing, cars, and electronics, each with assigned prices. The approach separates negotiation strategy from response generation, allowing for more controlled and flexible models. This uses real-world inspired products for negotiation, similar to our GUIDE dataset’s everyday scenarios, but focuses on buy-sell dynamics

rather than guiding intents [10].

The paper motivates decoupling by noting that end-to-end models make it hard to control the agent’s decision-making policy (in the sense of choosing good actions given a dialogue state), and reinforcement learning often leads to degenerate solutions like repetitive or unnatural language. To address this, they introduce a modular framework using coarse dialogue acts, such as *propose*(price=50), to represent the agent’s decision policy. This representation can then be optimized flexibly via reinforcement learning or domain-specific rules without causing degeneracy. For surface realization, they use a retrieval-based method that maintains context-awareness and produces diverse, natural utterances, evaluated on both the Deal or No Deal game and their new Craigslist dataset.

2.5 MindCraft

The MindCraft dataset [11] features 100 two-player games in Minecraft with about 20 text-based dialogue turns per game. Players have asymmetric knowledge and partial views, collaborating to create goal materials. It includes belief state questions and viewpoint videos. This is cooperative gameplay, similar to our task completion targeted work, but focuses on game play rather than our emphasis on real life guidance.

Unlike Deal or No Deal [7] (used to train end-to-end systems), which learn a single model to handle both dialogue generation and the agent’s decision policy directly from interaction data without intermediate annotations, MindCraft is used to build a modular system. It serves as a benchmark for modular approaches in which components such as belief state tracking, action prediction, and response generation are handled separately, relying on annotated belief states and viewpoint information for evaluation. For example, datasets like MultiWOZ [9] and TEACH [6] similarly support modular task-oriented pipelines with explicit stages for state tracking or task execution, in contrast to fully integrated end-to-end training. CraigslistBargain [10] also exemplifies this by explicitly decoupling the agent’s decision policy from surface generation into distinct modules.

While MindCraft emphasizes collaborative problem-solving in a simulated game environment with asymmetric information, the dataset itself does not prescribe any particular modeling paradigm; typical studies decompose tasks (e.g., belief/state inference versus response generation) and report separate metrics such as belief-state accuracy. By contrast, our GUIDE dataset targets asymmetric interactions about household activities and provides fine-grained intent labels tailored to instructional, corrective, and confusion-related phenomena. (Modeling choices are discussed in Chapter 4.)

2.6 Analysis of Dialogue in Human-Human Collaboration in Minecraft

The Minecraft Dialogue Corpus (MDC) from Narayan-Chen et al. (2019) [8] and the Collaborative Garden Task Corpus from Ichikawa and Higashinaka (2022) [12] analyze human-human collaboration in Minecraft. MDC has 509 dialogues where an architect and builder construct structures. The Garden Task has 500 text-chat dialogues for creating gardens, originally in Japanese and

translated to English. These are similar to two people doing something new together in our setup, but we focus on guidance intents, not collaboration.

2.7 SIMMC 2.0

SIMMC 2.0 (2021) [13] is a dataset of 11,000 dialogues for shopping assistants in fashion and furniture domains. It uses simulated photorealistic scenes with human-paraphrased utterances, annotated with four dialogue act types and five activity types. This is similar to our use of real-world data for training task-oriented systems, but our target is to provide guidance and processing language intents, not shopping assistance.

2.8 TEACH

The TEACH dataset (2021) [6] contains 3,000 human-human interactive dialogues for household tasks in simulation. One participant is the commander with oracle info, and the other is the follower, with interleaved utterances and actions. TEACH-DA adds 18 dialogue act types. This environment setting is close to ours, but focuses on training AI to perform tasks (teaching), while we emphasize processing intents in such instruction or other type of guidance utterances.

2.9 Emotional Support Dialogues

Additional datasets include ESConv (Liu et al., 2021) [14] with 1,000 text-based emotional support dialogues across five problem types, using eight support strategies. EmpatheticDialogues (Rashkin et al., 2019) [15] has 24,800 dialogues for emotional support with user-selected problems and emotion labels. These focus on emotional aspects, differing from our task-oriented teaching focus.

Chapter 3

Dataset Description

3.1 Overview of the Dataset

Recognizing intents is one of the most challenging tasks in situated human-agent communication, as there are a vast number of different purposes for verbal interaction—corresponding to the diverse situations communicators encounter—and much of interpreting communicative intent relies on inference. In such settings, the situation itself serves as part of the communication, for example, through physical demonstrations of actions, locations, or objects. In contrast, interactions between humans and large language models (LLMs), such as ChatGPT, are primarily information-oriented rather than action-oriented, lacking a shared physical context. This work investigates how well an offline classifier performs in recognizing intents across different kinds of contexts. In an interactive dialogue system, intent recognition would instead be conducted online, potentially requiring a different classification approach that incorporates real-time feedback and evolving dialogue states.

To address these challenges in a controlled yet realistic setting, we developed a specialized dataset for this thesis, drawing on collaborations with Pennsylvania State University NLP researchers.

The dataset developed by Pennsylvania State University NLP researchers used in this thesis is a custom collection of conversational transcripts designed to capture contextual intent variations in virtual human-alien interaction scenarios based on VirtualHome [16]. It serves as the foundation for training and evaluating the BERT-based intent classifier, enabling the model to learn how sentences express different intents based on subtle contextual differences, such as errors, confusion, or procedural variations in everyday tasks. We aim to predict differences in intent across different contexts and conversational targets, especially in skill-teaching scenarios such as sports coaching or, in our case, teaching household tasks.

The dataset is organized by the scenario, each corresponding to a specific setup representing

an everyday human activity. These setups simulate a human guiding an alien (a fictional entity unfamiliar with human norms) how to perform tasks, drawing from common household or routine actions. Examples of setups include `using_dishwasher`, `using_stove`, `read_book`, `watch_tv`, `bathroom_cabinet`, `bedroom_cabinet`, `using_computer`, `using_microwave`, `using_oven`, `wash_glass`, and `others` (examples in Next Page). This structure ensures a diverse range of scenarios, promoting generalizability in intent recognition.

For each setup, there are four variants, labeled from 0 to 3, which introduce controlled variations to highlight different gaps in the alien’s knowledge of household activities, and correspondingly, different ways for the human to guide the alien in household tasks. These variants are defined as follows:

- **0:** Represents the basic version of the scenario, where the interaction proceeds smoothly without misunderstandings made by aliens.
- **1:** Similar to variant 0 but includes one “lost moment,” where the alien experiences a brief period of being “frozen” due to cognitive overload or confusion.
- **2:** Consists of a variant where the alien handles the “wrong object,” introducing an error in object selection to simulate lack of knowledge of alien to perform task.
- **3:** Consists of a variant where the alien goes to the “wrong location,” introducing a spatial or navigational error.

Depend on 0-3 situation aliens are in, the human will provide different guidance to help alien (e.g. Provide help to guide alien out confusion. Provide feedback to encourage alien.)

The folder structure reflects this organization, with each variant stored as a separate CSV files. All setups are listed on next page and all of them have described 4 variations

Each CSV file contains transcripts of dialogues between the human and the alien for a specific setup and variant. The files are structured with the following columns:

- **user:** Indicates the speaker, either “Human” or “Alien.”
- **index:** A numerical index for the utterance within the dialogue sequence.
- **text:** The actual spoken text or phrase.
- **intent:** The initially assigned intent label (which may be preliminary or noisy).
- **correct_intent:** The corrected intent label used for training (if empty, the **intent** field is correct intent label).

This format captures turn-based conversations, allowing for the analysis of intent progression and contextual dependencies within a dialogue.

3.1.1 Intent Labels

The dataset employs a comprehensive set of intent labels, organized into five main categories (plus an “Other” category) to classify the intents behind each utterance. These labels are tailored to distinguish between human and alien perspectives, reflecting their roles in the guiding-practicing dynamic. The categories and sub-labels are as follows:



(a) Bathroom Cabinet



(b) Bedroom Cabinet



(c) Read Book



(d) Use Computer



(e) Use Dishwasher



(f) Use Microwave



(g) Use Oven



(h) Use Stove



(i) Wash Glass



(j) Watch TV

Figure 3.1: 10 scenes in GUIDE Dataset

1. Inform

This category involves one party providing information to the other, with human intents typically guiding the execution of household activities (e.g., through inform on actions, objects, or locations) and alien intents conveying the learner's state of knowledge (e.g., confusion, understanding, or condition) or providing affirmative/negative responses, often as implicit requests for help or even non-verbally. It is also worth noticing the motivation behind human and alien use. Inform intent is different. Human use of "inform" is to provide information, and alien use of "inform" is to announce their status.

Human Intent:

- Inform_action: Provides detailed information about the action (one specific action such as pick up something).
- Inform_object: Provides detailed information about the object.
- Inform_location: Provides detailed information about the location.
- Inform_next_step: Informs the alien about the next step (more general task need to perform next).
- Yes-no_positive: Positive answers to yes/no questions.
- Yes-no_negative: Negative answers to yes/no questions.
- Encouragement: Phrases like "Sounds good" or "Good job."
- Inform_other: Gives the alien other relevant information.

Alien Intent:

- Inform_confused: Indicates confusion or lack of understanding (through verbal informing).
- Inform_understanding: Indicates understanding.
- Explain_confused: Explains inability to act due to confusion (add explanation along with inform confusion).
- Display_confusion: Displays inability to act due to confusion (display through physical clue).
- Inform_observation: Making a remark or observation about something related to the main topic (usually unrelated to current task).
- Inform_condition: Inform human about their current condition (usually related to current task).
- Yes-no_positive: Positive answers to yes/no questions.
- Yes-no_negative: Negative answers to yes/no questions.

2. Request Information

This category involves seeking specific information from the other party, with human intents often probing the alien's knowledge or suggesting reconsideration (e.g., asking about awareness of actions, objects, locations, or next steps, or verifying readiness and need for help) to guide though alien to complete task process, while alien intents are more direct requests for clarification (e.g., about the next step, action, location, or object), reflecting the learner's asymmetric knowledge gaps and need for explicit guidance.

Human Intent:

- Ask_know_location: Asks if the other party knows the location of something.
- Ask_know_action: Asks if the other party knows the correct action.
- Ask_know_object: Asks if the other party knows the correct object.
- Ask_know_next_step: Asks if the other party knows the next step.
- Ask_okay: Asks if the other party is okay (to continue).
- Ask_want_help: Asks if the other party would like some help (understanding how to continue) .
- Ask_need_info: Asks if the other party needs more information.
- Ask_sure_about_action: Ask/suggest to alien to reconsider the action just taken or about to take.
- Ask_sure_about_object: Ask/suggest to alien to reconsider the object just handled or about to handle.
- Ask_sure_about_location: Ask/suggest to alien to reconsider the location just moved to or about to move to.

Alien Intent:

- Ask_next_step: Asks for the next action.
- Ask_about_action: Asks about the action.
- Ask_about_location: Asks about the location.
- Ask_about_object: Asks about the object.

3. Request Action

This category involves requesting the other party to perform a particular action, though there are fewer such intents overall because, in this type of guidance situation (and many others), verbal communication more often focuses on exchanging information rather than issuing commands or requests for actions. Due to the differences in the two speakers' roles, the intents are quite distinct: human intents encourage progression in the task, while alien intents seek overarching support. In principle, the human could request the alien to take the next step (using the human intent "Ask_take_next_step"), but this setup requires the alien to perform all actions as part of learning and practicing human ways, making such requests rare or absent.

Human Intent:

- Ask_take_next_step: Requests proceeding or continuing with the next step.

Alien Intent:

- Ask_guidance: General request for guidance on how to proceed.

4. Casual Talk

Casual interactions not directly related to the main purpose. These intent labels are set to record alien's behavior that are second nature to human.

General Intent (Human or Alien):

- Casual_observation: Remark or observation not directly related to the main topic.
- Casual_question: Question not directly related to the main topic.
- Casual_response: Response to a casual observation or question.

5. Greetings/Closings/Openings (Human or Alien)

- Extend_greetings: Says hello to the other person.
- Extend_goodbyes: Says goodbye to each other.
- Ask_get_started: Asks if the other party wants to start.
- Inform_get_started: Informs that one is ready to get started.

6. Other

Use this label if none of the others fit, and specify the intent.

These intent labels provide a granular framework for annotating dialogues, enabling the dataset to support supervised learning tasks focused on context-aware intent classification. Through this work, we aim to contribute a practical tool for advancing context-aware AI, paving the way for more intuitive human-machine interactions.

3.2 Dataset Acquisition and Preprocessing

The dataset was acquired through a two-step process in collaboration with a professor. In the first step, simulated interactions between two agents (a human primarily provide help to an alien who is trying to complete every day task) were generated in the VirtualHome environment [16], a programmable simulator for household activities that allows for the creation of realistic action sequences and visual scenarios without audio. This step produced video-based setups depicting various everyday tasks, such as using a dishwasher or reading a book, across controlled variants to introduce contextual challenges like errors or confusion.

In the second step, verbal communication was added to each turn of these simulated interactions. Advanced undergraduates were provided with the video recordings or descriptions of the human-alien scenarios and tasked with generating transcripts based on their own interpretations and understandings of the scenes. To facilitate this, an interface was developed to display instructions and the videos of the simulated interactions to the students. Pairs of students worked together, with each playing the role of human or alien, who inserted text utterances while stepping through the simulated video. The resulting sequence of text utterances constitutes a transcript for the video interaction. This crowdsourced approach leveraged diverse perspectives to create rich, varied dialogues, simulating real-world variability in intent expression. As a result, we collected an average of 11.175 transcripts per setup, culminating in a total of 447 distinct transcripts across all setups and variants.

Upon reviewing the collected transcript data, we identified significant inconsistencies in the initial intent labeling. The labels assigned by the students did not always adhere well to our standardized schema, leading to variations in interpretation and application of the intent categories. This inconsistency posed challenges for model training, as it introduced noise that could obscure underlying patterns in contextual intent shifts.

To address the noisy intent labels, we undertook a rigorous preprocessing phase focused on data cleaning and label refinement. I manually reviewed the entire dataset, based on predefined intent labeling rules (as outlined in Subsection 3.1.1). This expert annotator corrected and standardized the intent labels for each utterance, ensuring consistency across all transcripts. In cases where the original `intent` column was noisy or incorrect, the verified labels were recorded in the `correct_intent` column to facilitate comparison of the original labels with the revised labels, and thus to measure the impact of reducing the noise.

The data-cleaning step not only enhanced the dataset’s quality but also improved its utility for machine learning applications. Preliminary experiments with a BERT-based intent classifier demonstrated the impact of these efforts: the model’s average accuracy increased from an initial 0.61 (using the raw, inconsistent labels) to 0.70 (using the cleaned, standardized labels). This improvement underscores the importance of label consistency in enabling the model to better capture and analyze the intrinsic patterns within the transcripts, such as intent variations driven by contextual errors or confusion.

3.2.1 Additional Preprocessing Considerations

While the primary focus was on label correction, other minor preprocessing steps were applied as needed, including:

- Removal of any duplicate or incomplete transcripts to maintain dataset integrity.
- Normalization of the intent field, correction of misspellings, and use of consistent formatting.
- Verification of dialogue sequences to ensure logical flow and completeness within each transcript.

These steps collectively prepared a high-quality dataset suitable for training robust intent classification models.

3.2.2 Challenges and Limitations of the Dataset

While the dataset provides a valuable foundation for studying contextual intent variations, it is not without limitations. One significant challenge is the wide variation in the quality of the collected transcripts. A subset of the transcripts is detailed and rich in information, incorporating nuanced interpretations of the scenarios that effectively capture subtle contextual shifts, such as confusion or errors—making them highly suitable for our purpose of training a context-aware intent classifier. However, the majority of transcripts are simplistic, often serving as direct reflections of the video content without introducing meaningful variations in dialogue structure, phrasing, or intent expression.

This lack of diversity results in redundant information across many entries, which may not contribute optimally to machine learning models. On one hand, such redundancy could potentially aid the model in identifying core patterns by reinforcing common intent structures. On the other hand, it limits the dataset’s ability to represent the complexity of real-world conversations, where natural variations in language, cultural nuances, and spontaneous interactions are prevalent. Consequently, models trained on this dataset may not fully capture optimal patterns for intent recognition in diverse, uncontrolled environments, potentially leading to reduced generalizability.

Other limitations include the dataset’s focus on household and routine tasks, which may introduce domain-specific biases and limit applicability to broader contexts (e.g., professional or outdoor scenarios). Additionally, the reliance on student annotators from a single classroom setting could introduce demographic biases, such as age or cultural homogeneity, affecting the diversity of perspectives.

In future work, these issues could be addressed by expanding the range of contributors involved in transcript generation. For instance, sourcing inputs from a more diverse pool of participants—such as individuals from varied age groups, cultural backgrounds, or professional experiences—would increase the variance in transcript quality and content. This approach would not only mitigate redundancy but also augment the overall dataset size, providing more robust data for training and enabling models to better approximate real-world conversational dynamics. We can also expand our data collection to real-world cases by recording video and audio to obtain an optimal real-world record.

Chapter 4

Experiments on Context Size

This chapter presents the experimental evaluation of the BERT-based intent classifier, with a focus on how varying context sizes influence model performance in recognizing intents within the human-alien dialogue dataset. The experiments build upon the cleaned dataset described in Chapter 2, aiming to assess the model’s ability to capture contextual dependencies—such as errors, confusion, or procedural variations—that affect intent classification. We begin with baseline experiments to establish general performance metrics and model selection, before delving into intent-specific analyses and parameter variations in subsequent sections.

All experiments were conducted using the Hugging Face Transformers library [1], with models fine-tuned in a GPU-enabled environment (e.g., via a university cluster). The dataset was split into training (80%), validation (10%), and test (10%) sets, ensuring balanced representation across setups and variants (refer to chapter 3.1). Evaluation metrics include overall accuracy, as well as precision, recall, and F1-score for finer-grained analysis in later sections. Hyperparameters (Base model, Learning rate, Training epochs, etc.) were tuned empirically, with results averaged over three runs to account for randomness.

4.1 General Performance and Model Selection

Before turning to the intent classification experiments, we first evaluated the impact of data cleaning and compared several pre-trained language models. This section details the parameter tuning process, the effect of dataset preprocessing, and the selection of the optimal model based on overall accuracy. These baselines provide a foundation for subsequent experiments on context size variations.

4.1.1 Impact of Data Cleaning

As discussed in Chapter 2, the initial dataset contained noisy intent labels from student annotators, which were standardized through manual correction. To quantify the benefits of this preprocessing, we conducted comparative experiments using the same model and hyperparameters on both the uncleaned and cleaned versions of the dataset.

We used BERT-base-uncased [5] as the backbone model, fine-tuned with the following configuration:

- Model: bert-base-uncased
- Evaluation strategy: Epoch-based
- Learning rate: 2×10^{-5}
- Per-device train batch size: 16
- Per-device evaluation batch size: 16
- Number of training epochs: 3
- Weight decay: 0.01

These parameters were selected based on standard practices for fine-tuning BERT on classification tasks: a low learning rate to avoid catastrophic forgetting, moderate batch sizes for efficient training on limited hardware, and a small number of epochs to prevent overfitting given the dataset size (23,195 total samples across all transcripts).

On the uncleaned dataset, the model achieved an overall accuracy of 0.64. After applying the cleaned labels, accuracy improved to 0.70, representing a gain of 0.06. This improvement highlights the critical role of label consistency in enabling the model to learn meaningful patterns from the dialogues, such as intent shifts due to contextual errors (e.g., label makers didn't understand label meaning fully. Example on 4.1). Without cleaning, noise in the labels obscured these patterns, leading to suboptimal performance. All subsequent experiments in this chapter use the cleaned dataset to build on this enhanced baseline.

user	text	initial intent	correct intent
Human	We just need to put two of the towels in the cabinet.	Inform_action	Inform_next_step

Table 4.1: Example of misunderstood label

4.1.2 Comparison of Pre-trained Models

With the cleaned dataset as the foundation, we explored several pre-trained language models (PLMs) to identify the most effective model for our intent classification task. The goal was to balance performance, and computational efficiency. We tested the following models, all fine-tuned with the same hyperparameters as above (adjusted only for model-specific requirements, such as input formats for T5):

- **BERT-base-uncased** : A bidirectional transformer with 110M parameters, known for strong performance on sequence classification.
- **ALBERT-base** : A lighter variant of BERT (12M parameters) with parameter-sharing for efficiency.
- **T5-base** : A text-to-text transformer (220M parameters) adapted for classification by framing intents as generated text.
- **BERT-large-uncased** : A larger version of BERT (340M parameters) for potentially deeper contextual representations.

Results are summarized in Table 4.2, based on overall accuracy on the test set (averaged over three runs).

Table 4.2: Accuracy Comparison across models, and two settings for training epochs

Model	Accuracy: 3 Epochs	Accuracy: <3 epoch
ALBERT-base	0.67	N/A
T5-base	0.69	0.69 (E2)
BERT-base-uncased	0.70	0.72(E1)
BERT-large-uncased	0.70	0.71(E1)

BERT-base and BERT-large both achieved the highest accuracy of 0.70, outperforming ALBERT (0.67) and T5-base (0.69). While the differences are not statistically significant ($p > 0.05$ via paired t-tests across runs), BERT models demonstrated a slight edge in capturing the nuanced, context-dependent intents in our dataset—likely due to their bidirectional encoding, which aligns well with dialogue sequences involving back-and-forth interactions. ALBERT’s lower performance may stem from its reduced capacity, making it less effective for our moderately sized dataset, while T5’s text-generation paradigm required additional adaptations that introduced minor overhead without meaningful gains over ALBERT-base, or on performance with the BERT models..

Based on these results, we selected BERT-base-uncased as the primary model for further experiments. It offers a favorable trade-off between performance and efficiency (e.g., faster training times compared to BERT-large), making it suitable for iterative testing on context size variations. In the following sections, we analyze performance at the intent level and explore how hyperparameters and other conditions, such as context window size, affect outcomes (Table 4.2).

4.1.3 Hyperparameter Tuning: Training Epochs

Building on the baseline performance established with BERT-base-uncased at 0.70 accuracy using 3 epochs, we conducted experiments to explore the impact of varying the number of training epochs. The goal was to identify the optimal epoch setting that maximizes accuracy without overfitting, given the dataset size of 23,195 samples. We tested lower epoch counts to assess if shorter training could yield better or comparable results, potentially due to the model’s ability to quickly learn key patterns from the cleaned data. All other hyperparameters remained consistent with the

baseline (e.g., learning rate of 2×10^{-5} , batch size of 16, weight decay of 0.01), and results will be only affected by epoch number.

We started with BERT-large-uncased, reducing epochs to 1. This configuration achieved an overall accuracy of 0.72, surpassing the baseline of 0.70. The improvement suggests that BERT-large, with its deeper architecture, can capture contextual intent variations (e.g., shifts due to "wrong object" or "lost moment" scenarios) more effectively in fewer epochs, possibly avoiding overfitting on redundant transcript patterns.

Next, we tested BERT-base-uncased with 1 epoch, resulting in an accuracy of 0.71. This is a slight improvement over the 3-epoch baseline (0.70), indicating that for the base model, minimal fine-tuning is sufficient to learn essential dialogue features like intent progression in human-alien interactions. The gain may stem from the dataset's focus on household tasks, where patterns are repetitive enough to be grasped quickly.

For comparison, we also evaluated T5-base with 2 epochs, which yielded an accuracy of 0.69. This is lower than both BERT variants and aligns with earlier findings that T5 is less optimal for this classification task, even with adjusted epochs. The text-to-text framing of T5 may introduce unnecessary complexity for intent labeling, leading to suboptimal performance despite the epoch reduction.

These results highlight that lower epochs (e.g., 1) can enhance performance for BERT models, likely by preventing the model from memorizing noise in simpler transcripts while still leveraging the cleaned labels for context-aware learning. Based on this, we recommend 1 epoch for BERT-large as the best parameter for our setup, achieving the highest accuracy of 0.72. Future experiments could explore even finer epoch fractions or early stopping to refine this further.

4.1.4 Performance on High-Level Intent Categories

To further optimize the intent classification task, we explored aggregating the detailed intent labels into higher-level categories. This approach simplifies the classification problem by grouping similar intents, potentially improving model robustness and accuracy. The dataset's intents were reorganized into six broad categories: Casual, Greetings, Inform, Other, Request Action, and Request Information. These categories encompass the granular sub-labels (e.g., various "Inform" subtypes under the main "Inform" category), reducing the total number of classes and allowing the model to focus on broader patterns rather than fine distinctions.

This high-level categorization leads to higher accuracy due to its simplicity and broader range of information covered by each category. With fewer classes, the model has more data on each class, as subtle misclassifications within a category are less penalized, and the wider scope helps in capturing general contextual cues like overall dialogue flow in human-alien interactions.

We evaluated several models on this setup using the cleaned dataset (23,195 samples), with the same train-validation-test splits as before. Hyperparameters were kept consistent with prior experiments, adjusting only for model-specific needs.

Starting with BERT-large-uncased fine-tuned for 1 epoch, the model achieved an overall accuracy of 0.87 on the test set. This represents a substantial improvement over the fine-grained baseline (0.70-0.72), demonstrating that the larger BERT variant excels at learning high-level patterns quickly, even with minimal training iterations. The single epoch prevents overfitting on the dataset's redundant elements while leveraging the model's depth for better generalization across

categories like "Inform" (which dominates household task instructions) and "Request Information" (common in confusion scenarios).

For comparison, we tested T5-base with 3 epochs, yielding an accuracy of 0.86. Similarly, RoBERTa-base with 3 epochs also reached 0.86. Both models perform well but slightly trail BERT-large, possibly because RoBERTa's training optimizations and T5's generative approach are less aligned with the simplified classification here, where bidirectional context utilized in BERT appears to prove more effective for broad categories.

These results confirm that high-level categories enhance performance, with BERT-large at 1 epoch emerging as the top configuration (0.87 accuracy). The increased margin of error from broader categories makes the model more tolerant to variations in transcripts, such as those from "wrong object" or "lost moment" variants, paving the way for more reliable intent recognition in practical applications. Future work could refine these categories further or combine them with context size variations for even better outcomes. Also we can try to improve classification method from current one layer classifier to a hierarchical model. First classifier will classify by a board category and second classifier to determine detailed intent.

Table 4.3: Per-Intent Metrics for High-Level Categories (BERT-large and T5)

Category	BERT-large (1 Epoch)			T5 (3 Epochs)		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Casual	0.63	0.45	0.52	0.55	0.41	0.47
Greetings	0.85	0.81	0.83	0.88	0.78	0.82
Inform	0.90	0.95	0.92	0.89	0.94	0.92
Other	0.33	0.06	0.10	0.47	0.07	0.12
Request Action	0.39	0.18	0.25	0.32	0.21	0.25
Request Information	0.88	0.93	0.91	0.86	0.92	0.89
Macro Avg.	0.66	0.56	0.59	0.66	0.55	0.58
Weighted Avg.	0.86	0.87	0.86	0.85	0.86	0.85
Overall Accuracy	0.87			0.86		

Category	RoBERTa (3 Epochs)		
	Prec.	Rec.	F1
Casual	0.55	0.40	0.47
Greetings	0.86	0.79	0.83
Inform	0.89	0.94	0.92
Other	0.62	0.08	0.14
Request Action	0.38	0.21	0.27
Request Information	0.85	0.92	0.89
Macro Avg.	0.69	0.56	0.58
Weighted Avg.	0.85	0.86	0.85
Overall Accuracy	0.86		

4.2 Exploration of Multi-Utterance Context

In previous experiments, we used only one utterance as input to the model for intent classification. This limited the model’s ability to capture contextual dependencies across dialogue turns. To address this, we explored incorporating additional context by including 2 or 3 preceding utterances as input. The goal is to provide more surrounding information, to test the hypothesis that there is information in prior utterances a model could exploit to make more correct decisions.

We concatenated the additional utterances with the target one, separated by special tokens, and fed them into the model. This approach builds on the cleaned dataset from Chapter 2. Hyperparameters remained consistent with baselines, except where noted.

4.2.1 Results with Two Utterances

We first tested with two utterances (the target plus the previous one). Using BERT-large-uncased with 3 epochs, we obtained an overall accuracy of 0.73 on the test set. Reducing to 1 epoch improved this to 0.74.

For comparison, T5-base with two utterances and 15 epochs achieved 0.70 accuracy. During this run, we removed mistakenly duplicated data. Removing these data did not affect classifier performance but provided a significantly improved efficiency in training it. This reduced the dataset to 4,638 samples. The two runs before this (BERT-large-uncased with 3 epochs and 1 epoch) contains redundant data.

Next, we run T5-base at 9 epochs on 4,638 supports gave 0.68. BERT-large-uncased with 3 epochs and 2 utterances reached 0.74.

4.2.2 Results with 3 Utterances

Extending to three utterances (target plus two previous), T5-base with 5 epochs on the newly cleaned dataset yielded 0.62 accuracy. This lower performance may indicate that too much context introduces noise, especially in shorter dialogues.

Overall, using a larger context window improved performance relative to single-utterance baselines (e.g., from 0.70–0.72 to 0.74). The benefit is most apparent for intent types that often co-occur or appear in sequence—such as “Inform” and “Request Information”, where preceding turns help disambiguate the current label.

In the next section, we explore variations in padding methods to further optimize input handling.

4.3 Exploration of Padding Methods

Building on the multi-utterance experiments, we next tested different padding methods to handle variable-length inputs. Padding ensures consistent sequence lengths for model processing, but the method can affect how context is represented. We compared end padding (adding padding tokens at the end of the sequence) and inline padding (inserting padding between utterances).

These tests used BERT-large-uncased with 5 epochs and 3 utterances as input, on the cleaned dataset (4,638 samples). Other hyperparameters matched previous baselines.

End padding gave a consistent overall accuracy of 0.75 on the test set. This method preserves the natural order of utterances, allowing the model to focus on contextual flow without disruption.

In contrast, inline padding resulted in accuracies ranging from 0.43 to 0.75 across runs, with large inconsistency. This variability likely stems from how inline padding interrupts the sequence, making it harder for the model to capture dependencies between utterances.

Overall, end padding outperformed inline padding, improving stability and accuracy for multi-utterance inputs. In the next section, we discuss further explorations or analyses.

4.4 Further optimization

4.4.1 Optimization of Training hyperparameters

After exploring multi-utterance contexts and padding methods, we optimized the training arguments to further improve performance. This involved adjusting hyperparameters for better convergence and stability during fine-tuning.

The new arguments used are:

```
eval_strategy="steps",
eval_steps=100,
learning_rate=1e-5,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
num_train_epochs=10,
weight_decay=0.01,
logging_dir="./logs",
logging_steps=30,
save_steps=100,
save_total_limit=2,
load_best_model_at_end=True,
metric_for_best_model="accuracy",
warmup_steps=500,
fp16=False, # Disabled for debugging
```

Compared to the old arguments:

```
evaluation_strategy="steps",
eval_steps=100,
learning_rate=2e-5,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
num_train_epochs=5,
weight_decay=0.01,
logging_dir="./logs",
logging_steps=30,
save_steps=10,
save_total_limit=2,
```

These modifications were made to enhance training efficiency and model generalization. Specifically:

- Lowering the learning rate from $2e-5$ to $1e-5$ allows for more gradual updates to the model weights, reducing the risk of overshooting optimal values and potentially leading to better convergence on complex tasks like intent classification.
- Increasing the number of epochs from 5 to 10 provides more training iterations, giving the model additional time to learn from the data, which can improve accuracy but may increase the risk of overfitting if not monitored.
- Adding `warmup_steps=500` introduces a gradual increase in the learning rate at the start of training, stabilizing early updates and helping prevent divergence, especially with transformer models like BERT.
- Enabling `load_best_model_at_end=True`, with `metric_for_best_model="accuracy"`, ensures the model saved is the one with the highest accuracy on the validation set, improving reliability by selecting the best checkpoint.
- Adjusting `save_steps` from 10 to 100 reduces the frequency of saving checkpoints, saving computational resources while still allowing periodic evaluations.
- Disabling `fp16` (mixed precision) was done for debugging to ensure numerical stability, though it may slow down training on compatible hardware.

These changes collectively aim to refine the learning process, promoting steadier improvements and higher final performance, as evidenced by the accuracy gain from 0.74 to 0.75.

In the next section, we analyze intent-specific performance metrics.

4.4.2 Exploration of Token Length and Role Encoding

To further enhance the model's ability to distinguish between human and alien speakers, we explored increasing the token length for role indicators and simplifying their encoding. This aims to provide clearer separation in the input, potentially improving intent classification by emphasizing speaker identity in the context.

These changes increase the role token length from 2 to 5 for more robust representation, use binary encoding (0/1) instead of text labels for simplicity, remove unnecessary prefixes to streamline input, and incorporate early stopping with a patience of 2 to halt training if validation performance does not improve, preventing overfitting.

We expect this to help the model better capture speaker-specific patterns, such as differences in "Inform" intents between human instructions and alien responses, leading to more accurate classifications. The result comes out from 0.74 on baseline to 0.75.

4.4.3 Token Summing for Multi-Utterance Representation

As a final optimization, we replaced the default token appending method with token summing for combining multi-utterance inputs. This approach aggregates token representations by summing their embeddings, potentially capturing denser contextual information without extending sequence length excessively.

Key changes in the preprocessing and training script include: - Computing maximum token length based on actual tokenized inputs, capped at 128, for more precise padding. - Adding attention masks during utterance vectorization to preserve focus on meaningful tokens. - Building

combined sequences with separate handling of input IDs and masks, then summing token representations instead of simple concatenation. - Incorporating binary role encoding ("0" for Human, "1" for Alien) with a max length of 5. - Adding an EarlyStoppingCallback with patience of 2 to prevent overfitting. - Adjusting the train batch size to 12 and disabling mixed precision (fp16=False) for stability.

Using BERT-large-uncased with 2 utterances and 10 epochs under the optimized arguments, this method achieved an overall accuracy of 0.76 on the test set. This improvement over the previous 0.75 suggests that token summing enhances the model's ability to integrate context more effectively, leading to better intent classification.

4.5 Summary

Table 4.4: Summary of Accuracy Across Key Configurations

Configuration	Model/Epochs	Accuracy
Uncleaned Data Baseline	BERT-base / 3	0.64
Cleaned Data Baseline	BERT-base / 3	0.70
Low Epochs (Fine-Grained)	BERT-large / 1	0.72
High-Level Categories	BERT-large / 1	0.87
Two-Utterance Context	BERT-large / 1	0.74
End Padding (3 Utterances)	BERT-large / 5	0.75
Token Summing + Optimizations	BERT-large / 10	0.76

This chapter evaluated a BERT-based intent classifier on the GUIDE dataset, with a primary focus on how context size and input optimizations affect performance in recognizing intents from asymmetric human-alien dialogues. Through experiments, we established baselines, quantified the impact of data cleaning, compared pre-trained models, and explored multi-utterance contexts, padding strategies, and training refinements.

Key findings include: (i) Label cleaning improved accuracy by 0.06 (from 0.64 to 0.70), underscoring the importance of consistent annotations for learning contextual patterns; (ii) BERT models (base and large) outperformed alternatives like ALBERT and T5, achieving up to 0.72 with minimal epochs; (iii) Aggregating intents into high-level categories boosted accuracy to 0.87, simplifying the task while preserving broad patterns; (iv) Short multi-utterance contexts (e.g., target + 1 or 2 prior utterance) enhanced disambiguation for sequential intents like "Inform" and "Request Information," both raising accuracy to 0.74; (v) End padding provided stable gains (0.75), while optimizations like compact role encoding and token summing yielded a peak fine-grained accuracy of 0.76.

These results demonstrate that local context and lightweight input engineering can significantly improve intent recognition in situated, instructional scenarios, addressing limitations in existing datasets (Chapter 2) and supporting the thesis objective of advancing context-aware classification (Chapter 1). However, challenges remain, such as handling sparse intents and longer contexts without noise. Future extensions could include per-intent metrics, hierarchical classification, or multimodal integration, as discussed in the thesis conclusion.

Chapter 5

Conclusion

5.1 Summary of Contributions

This thesis investigated intent recognition in situated, goal-directed dialogues, focusing on asymmetric guidance interactions between a human guide and a novice “alien” learner in virtual household scenarios. The contributions are threefold:

- **Dataset:** We developed GUIDE (Growing Understanding through Interactive Daily Experience), a specialized dataset of 447 transcripts grounded in VirtualHome-based [16] simulations of 10 everyday activities, each with four controlled variants to elicit distinct forms of guidance and misunderstanding. The dataset includes fine-grained, role-sensitive intent labels that differentiate human and alien perspectives in guidance and support contexts.
- **Curation and Preprocessing:** We designed and applied a rigorous cleaning protocol to standardize intent labels across utterances. Manual correction reduced label noise and improved training reliability, yielding an absolute accuracy gain of approximately 0.06 (0.70; a 9% increase in accuracy, and a 17% reduction in error) under matched configurations.
- **Empirical Evaluation:** We conducted a systematic study of context size, model families, and training procedures for single-stage supervised intent classification. We quantified how context windows, padding strategies, tokenization choices, and aggregation methods affect performance on fine-grained and high-level intent categories.

5.2 Key Findings

We summarize the principal empirical results reported in Chapter 4:

- **Effect of Label Cleaning:** Under the same hyperparameters with `bert-base-uncased`, accuracy increased from the mid-0.60s on uncleaned data (e.g., 0.64) to 0.70 on the cleaned dataset, demonstrating that consistent labels are foundational for learning contextual intent patterns.
- **Model Selection:** Across tested pre-trained models, ALBERT-base (0.67) and T5-base (0.69) trailed BERT-base and BERT-large (both 0.70) on fine-grained labels. Given our limited sample size, we can't conclude which of these will be better option.
- **Context Size:** Incorporating prior turns improved performance relative to single-utterance inputs. With two-utterance context, accuracy increased to 0.74. Gains were most apparent for intents that tend to co-occur or occur in sequence (e.g., *Inform* followed by *Request Information*), where preceding turns help disambiguate the current label. Adding more context (three turns) did not uniformly help and, in some settings, introduced noise.
- **Input Handling:** End padding yielded an increase to 0.75 over inline padding, which produced unstable results (0.43–0.75). This supports preserving natural utterance order without inserting internal padding tokens.
- **Training Procedures:** Modest adjustments to learning rate, warmup, and epochs, combined with early stopping, yielded incremental gains (e.g., 0.75). Role-aware compact encodings and removal of non-essential prefixes shortened inputs and reduced token overhead without sacrificing speaker information.
- **Multi-Utterance Aggregation:** Replacing simple concatenation with token summing for multi-utterance representation bring accuracy to 0.76, suggesting denser contextual integration benefits this task under sequence-length constraints.
- **Coarser Categories:** Collapsing fine-grained intents into high-level categories substantially increased accuracy (up to 0.87 with BERT-large at 1 epoch), indicating the classifier captures broad communicative functions more readily than detailed subtypes.

5.3 Discussion

These findings collectively support three observations about intent recognition in situated, asymmetric guidance:

1. **Context Helps, but Economically:** Short, local context windows (e.g., the immediately preceding turn) deliver consistent benefits by exposing sequential regularities and implicit dependencies. Longer windows can dilute signal with irrelevant or redundant content in shorter dialogues.
2. **Quality over Quantity:** Data curation significantly impacts outcomes. The approximately six-point gain from label cleaning underscores that noise in intent supervision can obscure learnable contextual patterns more than architectural changes can recover.

3. Representation Matters: Lightweight choices in input construction—role encoding, padding strategy, and turn aggregation—yield material gains without increasing model size, highlighting the importance of engineering for context under finite sequence budgets.

Importantly, the approach here is a single-stage supervised intent classifier used as a component for recognition; it does not perform dialogue management, state tracking, or action generation. This distinction separates our modeling choices from full interactive agents and keeps the dataset discussion orthogonal to system design.

5.4 Limitations and Threats to Validity

Despite encouraging results, several limitations remain:

- Dataset Breadth and Quality: Transcripts vary in richness; many are near-descriptions of the scene with limited variation, which can reduce the diversity of intent realizations. Household tasks and a single academic cohort may introduce domain and demographic biases.
- Offline Classification: All experiments evaluate offline classification. In interactive settings, online intent recognition would need to incorporate evolving dialogue state, turn-taking dynamics, and potentially multimodal signals from the environment.
- Label Granularity and Balance: Some fine-grained intent subtypes are sparsely represented, which can depress per-class F1 and obscure true generalization. While high-level categories are easier to learn, they mask distinctions that matter for downstream behaviors.
- External Validity: Results are specific to the GUIDE label schema and VirtualHome-based scenarios. Generalization to other domains (e.g., professional workflows) or different embodiments (e.g., robotics) warrants further study.

5.5 Future Work

Building on this foundation, several directions appear promising:

- Data Expansion and Diversity: Collect transcripts from a broader pool of participants and settings, and incorporate audio alongside video to capture prosodic cues. Extend beyond household tasks to varied domains and cultures; quantify inter-annotator agreement to assess label reliability.
- Multimodal and Role-Aware Modeling: Integrate visual context (frames or program traces from VirtualHome) with text inputs to resolve referential ambiguity. Explore role-aware encoders or speaker-conditional adapters tailored to asymmetric knowledge states.
- Hierarchical and Online Classification: Implement hierarchical classifiers (coarse-to-fine) to mirror the strong performance of high-level categories while preserving fine-grained distinctions. Investigate online, incremental intent inference suitable for interactive systems with real-time feedback.

- **Robustness and Uncertainty:** Address class imbalance and label noise with robust losses and calibration; report per-intent precision/recall/F1 and confusion matrices; include statistical testing and confidence intervals across multiple runs.
- **Comparative and Cross-Domain Evaluation:** Benchmark on related situated datasets (e.g., TEACH) and evaluate transfer to non-household domains. Conduct ablation studies on context length, padding, and aggregation to determine minimal configurations with maximal generalization.

5.6 Concluding Remarks

This thesis introduced a curated dataset and a set of modeling and preprocessing practices for intent recognition in asymmetric, situated instructional dialogues. We demonstrated that (i) careful label standardization materially improves outcomes, (ii) compact, role-aware input design and short multi-turn context meaningfully aid disambiguation, and (iii) modest aggregation strategies can further lift performance under sequence constraints. Together, these results provide practical guidance for building intent recognizers that operate closer to real instructional settings and lay the groundwork for future, interactive agents that reason over both linguistic and situational context.

Bibliography

- [1] Hugging Face. Bert, 2018. Accessed: 2025-10-29.
- [2] Meta Fundamental AI Research Diplomacy Team (FAIR)[†], Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of *ipl*diplomacy*i* by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.
- [3] Yichi Zhang, Jianing Yang, Keunwoo Yu, Yinpei Dai, Shane Storks, Yuwei Bao, Jiayi Pan, Nikhil Devraj, Ziqiao Ma, and Joyce Chai. Seagull: An embodied agent for instruction following through situated dialog. In *Alexa Prize SimBot Challenge Proceedings*, 2023.
- [4] Prashanth Gurunath Shivakumar, Mu Yang, and Panayiotis Georgiou. Spoken language intent detection using confusion2vec. *arXiv preprint arXiv:1904.03576*, 2019.
- [5] Hugging Face. Bert base uncased, 2018. Accessed: 2025-09-26.
- [6] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2017–2025, Jun. 2022.
- [7] Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [8] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in Minecraft. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy, July 2019. Association for Computational Linguistics.
- [9] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz

- dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [10] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [11] Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [12] Takuma Ichikawa and Ryuichiro Higashinaka. Analysis of dialogue in human-human collaboration in Minecraft. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declercq, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4051–4059, Marseille, France, June 2022. European Language Resources Association.
- [13] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [14] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online, August 2021. Association for Computational Linguistics.
- [15] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.
- [16] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs, 2018.