Project Proposal: An NBA Fan AI Agent – Intent Classification and Slot Filling for NBA Statistics Queries

Kevin Zhou
yrzhou@umich.edu

Jiadong Zhu
jiadongz@umich.edu

1 Project Goals

As NBA fans, we often find ourselves spending significant time searching online for game information and player statistics. Older voice assistants like Siri were limited to basic queries such as scores and match times, and could not return detailed data like season averages or individual player stats. Newer systems such as Apple Intelligence and ChatGPT provide richer answers, but they rely on retrieving and synthesizing information from websites, which makes responses slower and less seamless. For fans who want quick, accurate, and user-friendly access to NBA-related data, these solutions are still not ideal.

This project aims to build a domain-specific NBA Fan AI Agent that delivers fast and reliable answers to basketball-related queries through natural language understanding (NLU), specifically intent classification and slot filling. Our initial plan is to construct a custom dataset of utterances across different intent categories and implement three approaches for comparison: Google Dialogflow ES, the OpenAI API, and fine-tuned machine learning models. If time permits, we will also integrate an NBA data query API to connect the NLU component with live information, effectively creating a real AI agent capable of answering questions. As an additional enhancement, we may include a speech-to-text module to support spoken interaction, making the system more powerful and user-friendly.

2 NLP Task Definition

2.1 Natural language understanding

The core natural language processing (NLP) task in this project is natural language understanding (NLU), which is the combination of intent classification and slot filling. Intent classification maps a user's query to one of several predefined categories, such as *game schedule*, *score lookup*, *player statistics*. Slot filling extracts structured information from the query, such as player names, team names, dates, times, and locations. For example, the query "What is Stephen Curry's average points per game this season?" should be classified under the *player statistics* intent, while the slots {player: "Stephen Curry", stat: "points per game", timeframe: "this season"} are extracted.

2.2 Speech-to-text (secondary)

As a secondary NLP component, we may also incorporate speech-to-text functionality. This would allow users to interact with the NBA Fan AI Agent using voice input rather than typed queries. The speech recognition module would transcribe the spoken query into text, which would then be processed by the NLU pipeline. While not the main focus of our project, this feature would enhance usability and bring the system closer to a real conversational agent experience.

3 Data

Since there are no existing public datasets tailored specifically for NBA-related conversational agents, we plan to construct our own dataset for intent classification and slot filling. We will initially define three intent categories: *game schedule*, *score lookup*, and *player statistics*. As the project progresses, we will expand this set to include two or three additional intents, aiming for a total of five to six categories. Each example will consist of a natural language utterance paired with its intent label and annotated slots such as team, player, datetime, and location.

We aim to build a dataset of roughly 600–800 utterances, with about 100–120 examples per intent class. Data will be created through a combination of (1) manual authoring by team members, (2) paraphrase generation using large language models followed by human validation, and (3) simple slot-value substitution such as replacing team or player names to expand coverage. For evaluation, we will split the dataset into training, development, and test sets, with additional out-of-domain queries included in the dev/test sets.

In designing our dataset, we will draw inspiration from prior NLU datasets such as ATIS (Hemphill et al., 1990), Snips (Coucke et al., 2018), and CLINC150 (Larson et al., 2019). While our domain is different, these resources provide useful guidance on structuring intent classes, defining slot schemas, and balancing examples across categories.

Here are a few preliminary examples we have drafted for the current three intents:

- Query: "When do the Lakers play next?"
  Intent: *game schedule*
  Slots: {team: Lakers}

- Query: "What was the score of yesterday's Pistons game?"
  Intent: *score lookup*
  Slots: {team: Pistons, datetime: yesterday}

- Query: "What is Stephen Curry's average points per game this season?"
  Intent: *player statistics*
  Slots: {player: Stephen Curry, stat: points per game, timeframe: this season}

This dataset will serve as the foundation for training and evaluating our NLU models. If time permits, we will also explore connecting the system to an NBA data query API to return live statistics, which would transform the dataset into a training corpus for mapping natural language queries to real API calls.

4 Related Work

4.1 Natural language understanding

The tasks of intent classification and slot filling have been extensively studied in the literature on natural language understanding (NLU), particularly within the context of spoken dialogue systems. Early work often treated the two tasks separately, but more recent approaches emphasize joint modeling for improved performance. Here, we highlight several representative papers that inform our project.

Benchmark datasets such as ATIS, Snips, and CLINC150 have become the standard for evaluating intent classification and slot filling models. ATIS focuses on airline travel queries and remains a long-standing benchmark, Snips covers diverse consumer domains such as music and weather, and CLINC150 provides 150 intents across domains along with out-of-scope queries for OOD evaluation. These datasets can inform the design of our NBA-specific dataset by illustrating how to balance intent coverage, slot schemas, and robustness to domain variation.

Guo et al. (2014) introduced recursive neural networks (RecNNs) for joint intent classification and slot filling, showing gains over CRF baselines on the ATIS dataset. Their approach leveraged syntactic parse trees for compositional representations, but was limited to airline travel queries and did not consider domain adaptation or noisy user input.

Goo et al. (2018) proposed the slot-gated model, which explicitly links intent detection and slot filling through shared gating mechanisms. Evaluated on ATIS and Snips, it improved sentence-level semantic frame accuracy and highlighted the importance of cross-task interaction—an idea directly applicable to NBA queries where intents and slots are tightly connected.

Chen et al. (2019) demonstrated that fine-tuning BERT achieves state-of-the-art results for intent and slot tasks, particularly improving rare entity performance. However, their experiments were restricted to benchmark datasets; adapting such models to NBA-specific data motivates our dataset construction.

Finally, Qian and Yu (2019) explored meta-learning for domain adaptation in dialogue systems, enabling rapid adaptation with minimal labeled data. This inspires our plan to include out-of-domain (OOD) evaluation to test generalization beyond the NBA domain.

In summary, prior work shows that (1) joint modeling outperforms pipeline approaches, (2) cross-task interaction and pretrained models boost performance, and (3) domain adaptation is key for specialized settings. Our project applies these insights to NBA fan queries by building a custom dataset and evaluating multiple architectures, expecting new challenges around noisy phrasing, player/team variants, and OOD rejection.

4.2 Speech-to-text (secondary)

If time permits, we may also extend our system with speech-to-text capabilities. Recent ASR models such as wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2022) have demonstrated strong performance and robustness across domains and noise conditions. These advances make it feasible to add reliable speech input as an additional interface for our NBA Fan AI Agent.

5 Evaluation

We will evaluate our system along two main dimensions: NLU quality and system-level performance.

5.1 Natural Language Understanding

For natural language understanding, we will use the following metrics:

- Intent classification accuracy: overall percentage of correctly predicted intents.
- Macro-F1 score: to account for potential imbalance across intent categories.
- Slot filling F1 score: token-level and span-level evaluation of slot extraction.
- Out-of-domain (OOD) detection accuracy: ability to reject queries that do not belong to any predefined intent.

For system-level performance, we will consider:

- Latency: average response time (p50 and p95) from input to final prediction.
- Cost: estimated cost per 1000 queries, which will vary between Dialogflow ES, OpenAI API, and local models.
- Scalability and maintainability: effort required to add new intents or slots to each system.

5.2 Speech-to-text (secondary)

If we add speech input, we will separately evaluate the ASR component and its impact on the end-to-end system:

- Word Error Rate (WER) and Character Error Rate (CER): transcription accuracy on our basketball query set.
- Domain term accuracy: error rates specifically on player names, team names, and basketball statistics terms.
- Latency: p50 and p95 time from audio end to transcript availability.
- Robustness: WER under common noise conditions (indoor noise, crowd ambience) and varied speaking rates or accents.
- Endpointing accuracy: correctness of start/stop detection in streaming mode.
- End-to-end impact: change in intent accuracy and slot F1 when using ASR transcripts versus gold text, including cascading error analysis.

6 Work Plan

6.1 Timeline

- Weeks 1–2 (9/25–10/8): Finalize intent categories (initially *game schedule*, *score lookup*, *player statistics*, later expanding to 5–6). Define slot schema (team, player, datetime, stat, location, etc). Begin dataset creation through manual authoring, paraphrase generation, and slot-value substitution. Each team member will generate data for half of the intents.
- Weeks 3–5 (10/9–10/29): Build baselines. Implement a simple classifier and configure Google Dialogflow ES. By October 29, have a functioning baseline system ready for the project update deadline.
- Weeks 6–10 (10/30–12/3): Implement advanced systems. One team member will research, train, and evaluate machine learning models. The other team member will implement and evaluate an OpenAI API few-shot approach. Both will then conduct system-level evaluation. If time allows, explore optional extensions including connecting to an NBA stats API and adding speech-to-text.
- Weeks 11–12 (12/4–12/15): Final integration and reporting. Summarize evaluation results, conduct error analysis, and prepare the poster presentation by December 8. Finalize the written report by December 15.

6.2 Two-person team justification

This project requires two contributors because it involves dataset construction, multiple NLU approaches, and system-level evaluation. The division of work is as follows:

- Team member 1: Responsible for half of the dataset creation, configuration of Dialogflow ES, and implementation of the OpenAI API few-shot approach.
- Team member 2: Responsible for the other half of the dataset creation, implementation of baseline classifiers, and training and evaluation of transformer-based models.

Both team members will collaborate on dataset validation, evaluation metrics (intent accuracy, macro-F1, slot F1, OOD detection, latency, and cost), error analysis, poster preparation, and the final report. This ensures balanced contributions to the core NLP tasks, while still allowing complementary specialization in different approaches.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *Preprint*, arXiv:1902.10909.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *Preprint*, arXiv:1805.10190.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 554–559.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, page 96–101, USA. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. *Preprint*, arXiv:1906.03520.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.