FIREBALL!

A Dataset of Dungeons and Dragons Actual-Play with Structured Game State Information

Andrew Zhu, Karmanya Aggarwal, Alexander Feng
Lara J. Martin, Chris Callison Burch
{andrz, karmanya, ahfeng, ccb}@seas.upenn.edu; laramar@umbc.edu

Penn
University of Pennsylvania

FIREBALL! is the largest dataset containing structured game states from real Dungeons and Dragons games, capturing **8 million** utterances and **1.3 million** game states.

INTRODUCTION

- D&D is a language-based role-playing game
- Parts of game have concrete state, e.g. fights w/ monsters
- Structured game state in NLG tasks lead to higher quality generations
- Existing datasets lack game state or rely on post-hoc state prediction
- Applications like Avrae can be used to track game state during gameplay

TASKS

- We show how FIREBALL can be used to improve performance on NLG tasks
- 2 tasks: Utterance to Command and State to Narration
- Compare LLM (GPT-3) performance with and without FIREBALL

Utterance to Command

STATE:
Actors:
- Filgo Bitterfoot (Mountain Dwarf; Fighter 5) <43/43 HP; Healthy>
- BU1 (Bulette) <42/53 HP; Injured>

Current:
Name: Filgo Bitterfoot
Class: Fighter 5
Race: Mountain Dwarf
Attacks: Greataxe, Longsword, Longbow
Actions: Second Wind, Action Surge

UTTERANCE: Filgo swings his axe at the bulette! "Raaaargh!"

COMMAND: ***!attack greataxe -t bu1***

State to Narration

STATE:
Actors:
- Khaslar (Reborn; Blood Hunter 5) <23/49 HP; Bloodied> [Feeling Inspired (1d6)]
- GFoY1 (Gnoll Fang of Yeenoghu) <12/65 HP; Injured> [Marked (Faerie Fire)]

Current:
Name: Khaslar
Class: Blood Hunter 5
Race: Reborn
Attacks: Crossbow, Light, Fire Talon
Effects: Feeling Inspired (1d6)

ACTION: Khaslar attacks GFoY1 with a Fire Talon and hit! GFoY1 took 14 damage.

GENERATION: ***Thanks to the inspiration and the faerie fire, he guts the gnoll with fire, blade and lightning.***

DISCUSSION

- FIREBALL models correctly match player intent when predicting commands
- We tested generated commands by running them in the Avrae system
- FIREBALL models generate cohesive, grounded narration compared to baseline models, evaluated by 45 D&D players
- Narrative generations should preserve player agency
- LLMs can have poor numerical reasoning (e.g. HP) - reporting bias
- These models can be used to assist new players and inspire DMs
- We're excited to see how future work utilizes our dataset!

DATASET STATISTICS

Utterances	8,012,706
Commands	2,109,603
Combat States	1,297,254
Unique Actors	161,501
Utterance to Command Examples	120,397
State to Narration Examples	43,372

UTTERANCE TO COMMAND

(Automated Evaluation)

	Pass Rate	Unit Tests	Avg SGleu	Avg RougeL
FIREBALL -FULL	**0.726**	**0.65**	**0.355**	0.75
FIREBALL -NOSTATE	0.235	0.234	0.189	0.551
FEWSHOT -FULL	0.432	0.429	0.325	**0.771**
FEWSHOT -NOSTATE	0.319	0.25	0.246	0.598

STATE TO NARRATION

(Automated + Human Evaluation)

	Sense	Specific	Interest	BERT-Score	Perplexity
FIREBALL -FULL	**55.48**	47.1	4.6	0.848	208.989
FIREBALL -SHORT	51.61	**48.39**	**4.98**	0.848	202.398
COMMAND	40.97	37.1	4.72	0.842	**156.984**
DIALOG	35.81	27.42	4.27	0.846	208.968
GOLD (Human)	53.55	47.74	4.91	N/A	452.653

Download the paper

Download the dataset