Bachelor Project

Authors: Seyed Ali Hosseininasab, Farima Kafi

Instructor: **Maryam Tahmasbi**

Date Last Edited: January 23, 2026

1 Introduction

1.1 Bridging the Gap between Expert and Language Models

Kim et al. [2] address a fundamental challenge in explainable artificial intelligence: bridging the gap between expert decision-making models and large language models (LLMs). While expert models such as modern chess engines achieve superhuman performance in complex decision-making tasks, their outputs are difficult for humans to interpret. In contrast, LLMs are capable of generating fluent and natural language explanations but often suffer from hallucinations and lack domain-specific reasoning capabilities. The authors study this problem in the context of chess commentary generation, which serves as a representative task for explaining complex decisions through natural language.

To overcome these limitations, the paper proposes a framework called *Concept-guided Chess Commentary* (CCC). The key idea is to extract high-level chess concepts—such as king safety, passed pawns, mobility, and threats—from a neural chess expert model using concept-based explanation techniques. These concepts are represented as vectors learned via linear classifiers trained on positions labeled by a classical chess engine. For a given move, the framework prioritizes concepts by comparing their relevance before and after the move, identifying which concepts are most influential in explaining the decision.

The prioritized concepts are then provided as guidance to a large language model during commentary generation. This integration allows the LLM to focus on strategically important aspects of the position while avoiding incorrect or hallucinated explanations. As a result, the generated commentary is both linguistically fluent and strategically accurate.

In addition to commentary generation, the authors address the problem of evaluation. They introduce *GPT-based Chess Commentary Evaluation* (GCC-Eval), a multi-dimensional evaluation framework that assesses commentary based on relevance, completeness, clarity, and fluency. Unlike traditional similarity-based metrics such as BLEU or ROUGE, GCC-Eval incorporates expert chess knowledge and shows a significantly higher correlation with human expert judgments.

Experimental results on a standard chess commentary dataset demonstrate that the proposed CCC framework outperforms existing baselines and even rivals human-generated commentary in terms of informativeness and linguistic quality. Overall, this work highlights the effectiveness of concept-based explanations as an interface between expert models and language models, and it provides a generalizable approach for interpretable decision explanation beyond the domain of chess.

1.2 Learning to Generate Move-by-Move Chess Commentary

Jhamtani et al. [1] study the task of automatically generating natural language commentary for individual moves in chess games. The authors frame this problem as a grounded natural language generation task, where the generated text must accurately reflect the game state, adhere to the rules of chess, and capture the strategic and pragmatic motivations behind a move. Chess commentary generation is particularly challenging due to the need for domain knowledge, the evolving nature of the game state, and the high variability in how humans describe the same move.

A major contribution of this work is the introduction of a large-scale chess commentary dataset collected from online chess forums. The dataset contains more than 298,000 move–commentary pairs spanning over 11,000 games, making it the first dataset of this scale for move-level chess commentary generation. An analysis of the data reveals substantial diversity in commentary styles, which the authors categorize into several types, including direct move descriptions, evaluations of move quality, comparisons with alternative moves, strategic planning explanations, contextual game-level information, and general comments. This diversity highlights the importance of content selection and pragmatic reasoning in the generation process.

To address these challenges, the authors propose a neural model called *Game-Aware Commentary* (GAC). The model is trained end-to-end and incorporates multiple sources of domain-specific information extracted from the chess board state before and after a move. These features include move-related information (e.g., piece type and position), threat-related information (e.g., attacking and defending pieces), and score-based features obtained from a chess engine to estimate move quality. Feature representations are embedded in

a shared continuous space, and a bidirectional LSTM encoder is used to model feature conjunctions. A selection mechanism is further employed to dynamically identify salient features during text generation, followed by an LSTM decoder that produces the commentary.

Experimental results show that the proposed GAC model outperforms several baselines, including template-based methods, nearest-neighbor approaches, and models that rely solely on raw board representations. Although automatic metrics such as BLEU yield relatively low scores due to the inherent variability of natural language commentary, human evaluation demonstrates that the generated commentaries are comparable to, and in some cases indistinguishable from, human-written commentary in terms of correctness and fluency. Overall, this work establishes a strong foundation for research on grounded and interpretable language generation in games and serves as a key reference for subsequent studies in chess commentary and explainable decision-making systems.

2 References

References

[1] Harsh Jhamtani, Varun Gangal, Eduard Hovy, Graham Neubig, and Taylor Berg-Kirkpatrick. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1661–1671, Melbourne, Australia, 2018. Association for Computational Linguistics.

[2] Jaechang Kim, Jinmin Goh, Inseok Hwang, Jaewoong Cho, and Jungseul Ok. Bridging the gap between expert and language models: Concept-guided chess commentary generation and evaluation. *arXiv preprint arXiv:2410.20811*, 2025.