Approach for Hangman Solver

Sumit Prabhakar

April 5, 2025

Overview

The Hangman Solver employs a layered and adaptive strategy to guess letters in a partially revealed word. It combines statistical frequency analysis, heuristic weighting, and fallback mechanisms, all designed to emulate intelligent human guessing with enhanced accuracy.

Approach

The solver's decision pipeline is divided into the following core components:

1. Dictionary Initialization and Preprocessing

- A corpus of 250,000 English words is loaded and sanitized.
- Substring dictionaries of sizes 3 to 30 are constructed for efficient matching.
- Vowel ratios are computed for all words to inform later heuristics.

2. Filtering Based on Word Pattern

Each guess round starts with filtering words matching the current partially-known pattern using regular expressions. This forms the candidate word pool for that step.

3. Position-Specific Frequency Analysis

- For unknown positions (underscores), letter frequencies at each position are calculated.
- Helps prioritize letters that statistically appear more often in those specific positions.

4. Weighted Letter Scoring

Each letter is assigned a composite score combining:

i. **Letter Frequency Score**: How frequently a letter appears in the filtered dictionary.

ii. **Positional Bonus**: Extra weight if it appears often in unknown positions.

iii. **Word-Based Weighting**:

- Words ending in common suffixes (e.g., `ing`, `ed`, `s`) receive a boost.
- Words with balanced vowel ratios (0.3 to 0.5) are considered more probable.

5. Heuristic Pruning

- Vowels are deprioritized if the current pattern already has a high vowel ratio (to avoid over-guessing them).
- Previously guessed letters are ignored.

6. Multi-Stage Fallback Mechanism

If the main frequency scoring yields no good candidates, the solver uses:

1. Substring-based frequency analysis on segments of the pattern.
2. Shorter n-grams extracted from the full dictionary to infer probable letters.
3. Global letter frequency distribution from the full dictionary as a last resort.

7. Adaptive Memory

Guessed letters are stored to avoid repetition and adapt future guesses. This keeps the guessing logic sharp and memory-aware.

Key Strengths

- **Multi-source intelligence**: Combines local (position-aware) and global (frequency-based) data.
- **Fallback resilience**: Always has a backup strategy when the optimal guess is ambiguous.
- **Data-driven intuition**: Mimics human guessing behavior using measurable patterns.

Conclusion

This solver demonstrates a thoughtful integration of statistics, pattern recognition, and heuristics. It maintains flexibility across word lengths and structures while minimizing the number of wrong guesses through predictive reasoning.