Source Coding for *the Game of Thrones*

2nd-Order Markov Adaptive Approximation (AME), Huffman, and Fano Coding

Jinming Ren (2022190908020, Presenter)
Yuhao Liu (2022190908022)

University of Glasgow
University of Electronic Science and Technology of China

January 12, 2025

Overview

Background

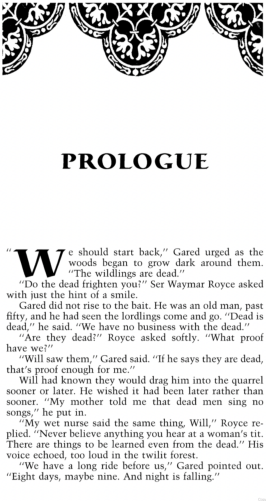
- **Goal:** Compress text (first 3 chapters of *the Game of Thrones*) effectively using a source coding method.

Theorem (Source Coding Theorem)

$$\bar{L} \geq H(X)$$

- **Entropy Coding:** What's beyond?
  - Huffman
  - Fano
  - Arithmic
  - ...
- "Memoryless Assumption"

```
I love yo_
```



PROLOGUE

"We should start back," Gared urged as the woods began to grow dark around them. "The wildlings are dead."

"Do the dead frighten you?" Ser Waymar Royce asked with just the hint of a smile.

Gared did not rise to the bait. He was an old man, past fifty, and he had seen the lordlings come and go. "Dead is dead," he said. "We have no business with the dead."

"Are they dead?" Royce asked softly. "What proof have we?"

"Will saw them," Gared said. "If he says they are dead, that's proof enough for me."

Will had known they would drag him into the quarrel sooner or later. He wished it had been later rather than sooner. "My mother told me that dead men sing no songs," he put in.

"My wet nurse said the same thing, Will," Royce replied. "Never believe anything you hear at a woman's tit. There are things to be learned even from the dead." His voice echoed, too loud in the twilit forest.

"We have a long ride before us," Gared pointed out. "Eight days, maybe nine. And night is falling."

Figure 1: *The Game of Thrones* by George R. R. Martin

Motivation

- **Non-entropy Coding:** Respect the internal structure of the source.
  - LZ family (LZ77, LZ78, LZW, LZMA, LZSS)
  - PPM (Prediction by Partial Matching)
  - DMC (Dynamic Markov Compression)
  - ML-based
  - ...
- Common point: "Prediction"!

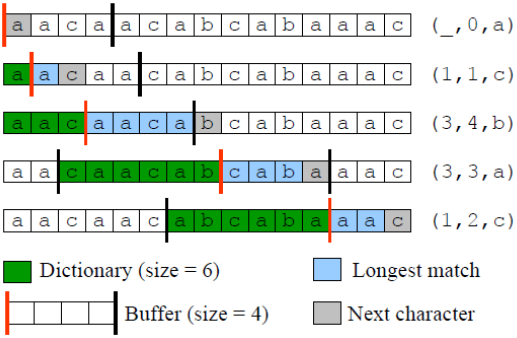


Figure 2: LZ77 Compression

The Future of Lossless Source Coding

Compression = Prediction

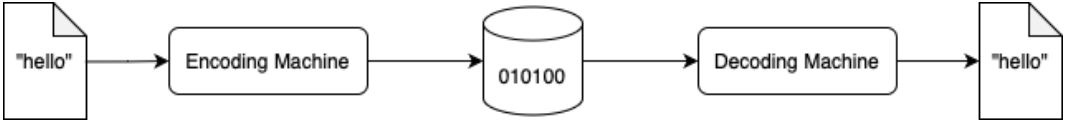


Figure 3: Same deterministic machines at the sender and the receiver

Claim 1

A good compressor is also a good predictor.

Claim 2

The bit stream from a good compressor should be unpredictable.

Implementation Flowcharts

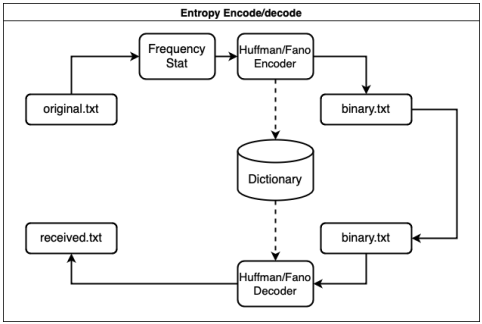


Figure 4: Without AME

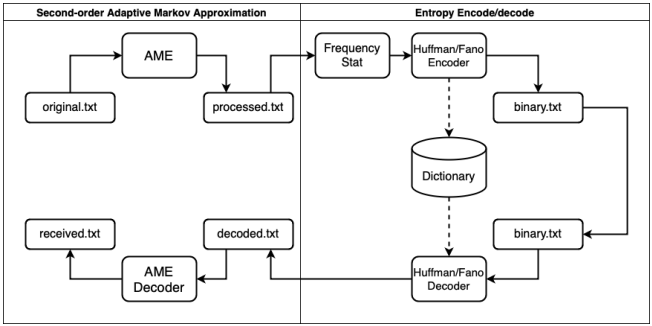


Figure 5: With AME

Inside AME Block

`catcatme`

`catc2me`

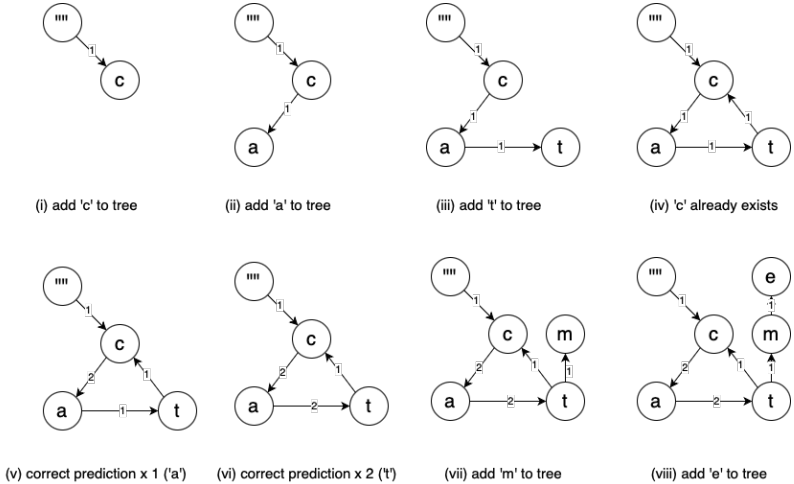


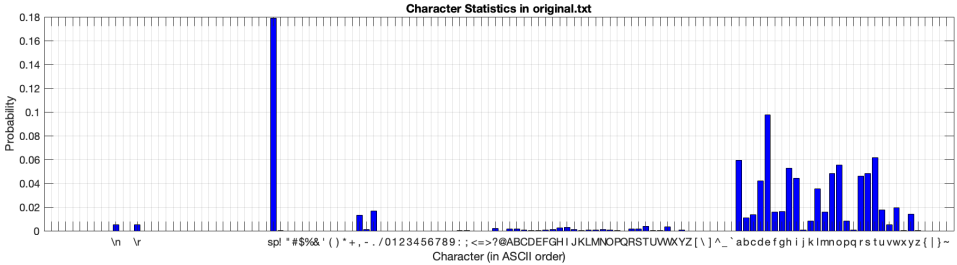
Figure 6: Example: Contruction of an AME tree

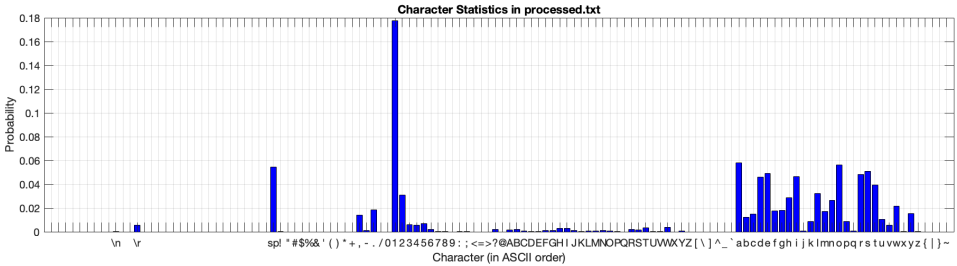
AME Processed Text

```
Ned squeezed her hand. There must be a feast, of course, with singers
    , and Robert will want to hunt. I shall send Jory south with an
    honor guard to meet them on the kingsroad and escort them back.
    Gods, how are we going to feed them all? On his way already, you
    said? Damn the man. Damn his royal hide.
```

```
1Ned1sq1eez1d1h1r ha1d.13r2must b2a feast,1of1co1rse,1wit1 si1gers,1
    a1d1R1b1rt wil1 w2t 1o hu1t.1I shal1 send1J1ry1so1t1 wit1 a2honor
     guard2o m1et 3m on5ki1gsroad1a1d1escort 3m back.1Gods,1how
    ar2we1goi1g2o feed4m al1? On1his1w1y1alr1ady,1yo1 said? Damn5ma1
    .1Damn1his1royal hide.
```

Frequency Statistics Comparison





Compression Performance

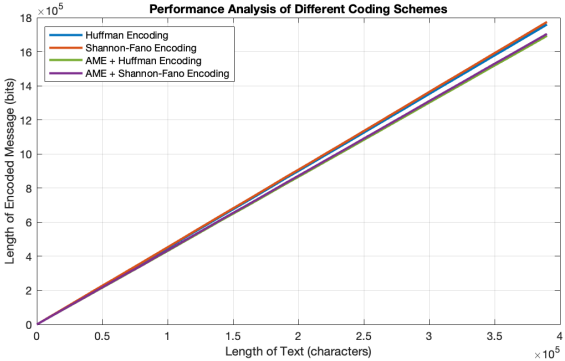


Figure 7: Performance Comparison

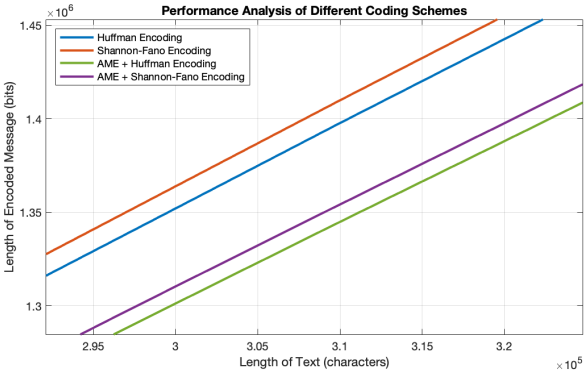


Figure 8: Zoomed-in View

Compression Performance (Cont'd)

Table 1: AME shortened the length of the binary file

Method	$\bar{L}$	**With AME**	**Without AME**	**Improved**
Huffman	4.5 b/char	207797 b	217014 b	5.25%
Fano	8.4 b/char	209343 b	218913 b	4.37%

Conclusion and Future Work

- **Huffman vs. Fano:** Huffman is typically better.
- **AME's Advantage:** Pre-processing text to capture the "memoryless" component of the source.
- **Overall Gains:** A modest but consistent improvement of $\approx 4 - 5\%$ in code length for this text.
- **Future Work:**
  - Higher-order Markov models.
  - Fast speed neural network models.
  - Breakthrough in Linguistics, uncover general pattern in language syntax.

References

C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379423, July 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Trans.* Inform. Theory, vol. 23, no. 3, pp. 337343, May 1977, doi: 10.1109/TIT.1977.1055714.

A. D. Wyner and J. Ziv, The sliding-window Lempel-Ziv algorithm is asymptotically optimal, *Proc. IEEE*, vol. 82, no. 6, pp. 872877, Jun. 1994, doi: 10.1109/5.286191.

T. Sharma et al., "A Survey on Machine Learning Techniques for Source Code Analysis," Sep. 13, 2022, arXiv: arXiv:2110.09610. doi: 10.48550/arXiv.2110.09610.

Advances in Communication and Computing Technologies (ICACACT 2014), Mumbai, India, 2014, pp. 1-6, doi: 10.1109/EIC.2015.7230711.

N. Dhawale, "Implementation of Huffman algorithm and study for optimization," 2014 International Conference on Advances in Communication and Computing Technologies (ICACACT 2014), Mumbai, India, 2014, pp. 1-6, doi: 10.1109/EIC.2015.7230711.

S. Congero and K. Zeger, "Competitive Advantage of Huffman and Shannon-Fano Codes," in *IEEE Transactions on Information Theory*, vol. 70, no. 11, pp. 7581-7598, Nov. 2024, doi: 10.1109/TIT.2024.3417010.