BOOKWORM: A Dataset for Character Description and Analysis

Argyrios Papoudakis, Mirella Lapata, Frank Keller
School of Informatics, University of Edinburgh

Paper & Code

Motivation

Characters are at the heart of every story!

NLP models should be able to understand characters if we want to summarize, analyze and generate stories.

Understanding characters of long narratives (~100k words) is challenging:

- the plot is complex
- multiple interacting characters (protagonists, antagonists, minor)
- length challenging for the current models

Tasks and Dataset

BOOKWORM

- **Two text-generation tasks:**

Character Description: short profile of a character (e.g., role, relationships, actions, attributes) (Brahman et al., 2021)

Character Analysis: in-depth interpretation of a character (e.g., personality, motives, development, greater level of detail)

Book: Bleak house, **Character**: Esther Summerson

Character Description

The narrator and protagonist. Esther, an orphan, becomes the housekeeper at Bleak House when she, Ada, and Richard are taken in by Mr. Jarndyce. Everyone loves Esther, who is selfless and nurturing, and she becomes the confidante of several young women. Although she eventually does find her mother, circumstances prevent them from developing a relationship. At first a hesitant, insecure narrator, Esther's confidence in her storytelling grows, and she controls the narrative skillfully.

Character Analysis

Esther Summerson, the narrator and protagonist of Bleak House, [..] she proves to be a confident narrator who never misses the opportunity to relate others' compliments of her.[..] As her narrative gains breadth and depth, her confidence as a narrator grows. She deliberately withholds information or delays including it to give her story coherence and dramatic effect. And even though she is for the most part a reliable narrator (a narrator we can trust to accurately tell the story), she is less reliable when relaying information about her romantic life. Esther nurtures everyone around her, and her first instinct is to be motherly, perhaps because she has never had a caring mother figure of her own. [..] Ironically, Esther, for all her caring and tenderness, is the unwitting cause of great unhappiness. [..] Because of Esther's illegitimate birth, Lady Dedlock was forever estranged from her sister, Miss Barbary, and was forced to carry a painful secret. Because other unhappinesses, [..] we could argue that Esther is indirectly responsible for these as well.

- **Joint Character Description**: generate a description for each character of a story sequentially.

Data Collection

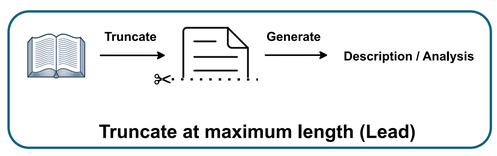
1. Book Stories from Gutenberg Project
2. Human-written Descriptions and Analyses from Study Guides Websites (e.g., SparkNotes, LitCharts)

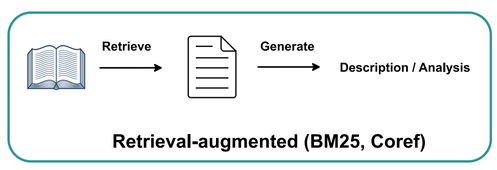
Dataset Statistics

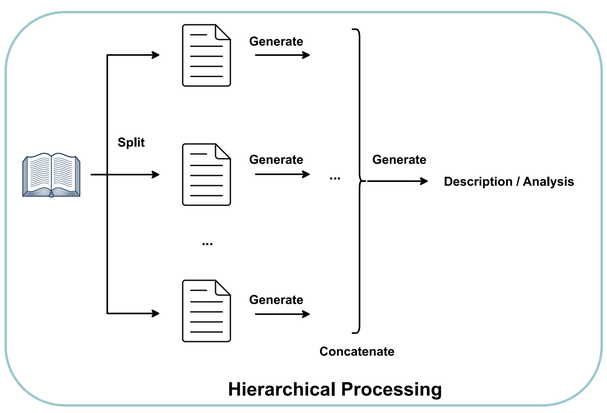
Dataset	Books	Samples	Avg. Input Length	Avg. Output Length
Description	324	5,869	97,685.82	88.79
Analysis	133	1,328	95,758.79	602.65

	Novel n-grams %		
	Unigrams	Bigrams	Trigrams
BOOKWORM	49.60	83.81	95.96

Experiments

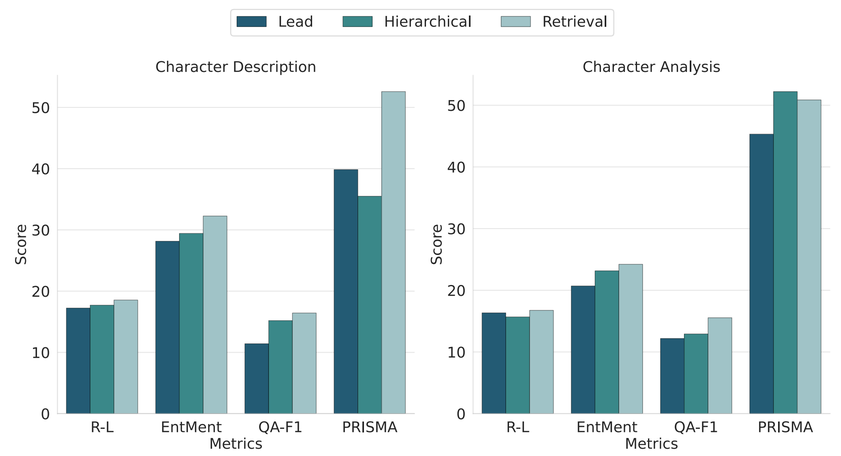




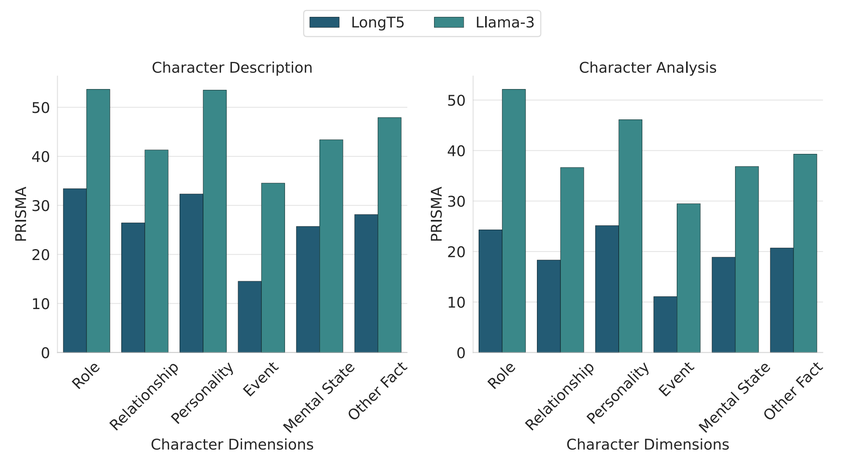


Results

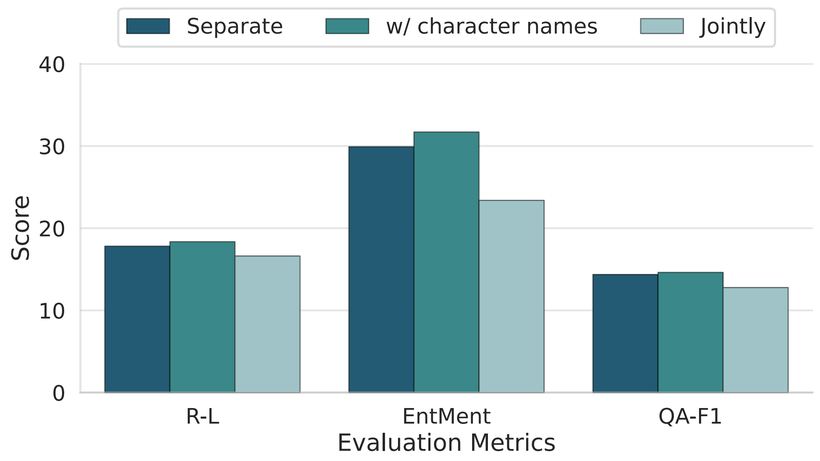
Retrieval-augmented approach performs best in both tasks.



Facts related to events and relationships are harder to get right.



Easier to describe one character than many.



Key Takeaways and Future Work

1. Retrieval performs best, which is in contrast with book-length summarization, where hierarchical method is the de facto approach.
2. Models struggle with dynamic aspects of characters (e.g., events, relationships) more than the static ones (e.g., role, personality).
3. Joint description of characters increases difficulty (more characters, longer output) and requires a better modeling.
4. There is a need for research on new stories (not publicly available) and efficient evaluation metrics which use the full input.