Detail in the Novel

Tim Schott
School of Information
University of California, Berkeley
timschott@berkeley.edu

1 Overview

Authors produce narratives by complementing events that advance a plot with details that evince deeper meaning from the work's characters, context and setting. Intriguing efforts to understand the functional structure and patterns of narratives are carried out through traditional modes of literary critique as well as through digital means of analysis. These treatments typically emphasize extracting the most consequential and salient aspects that comprise a story with the end goal of performing tasks like machine summarization. A narrative's details, conversely, often only garner attention during traditional close-readings that lack the capacity to situate a novelist's usage of details within a broader artistic, social, and political context.

In fact, outside of (eloquently-written) conjecture, literary scholars grievously lack a clear definition of details and their implications; when it comes to these particulars, they're anything but (Auerbach, 2003; Miles, 1979). The objective of this project is to analyze a corpus of English-language novels to produce heuristics for recording and rating their usage of detail. Furnishing a coherent framework such as this would induce progress towards a concrete understanding of the numinous multiplicity of details in the novel and empower future scholars in the computational humanities to investigate details on a large scale.

2 Related Literature

2.1 Narratology, Salience, Details

Literary critic Roland Barthes provides theoretical foundation for the field of narratology when he taxonomizes components of narrative (Barthes and Duisit, 1975). One of the elements he identifies is the "cardinal function." A cardinal function refers to an action that "directly [affects] the continuation of the story ... it either initiates or resolves an uncertainty." Barthes also narrows his gaze and grapples with details. In addition to cardinal functions, every narrative possesses a certain number of details (Barthes, 1989). He contrives a category called "informants" that produce "ready-made knowledge" and exist "to root fiction in the real world (Barthes and Duisit, 1975). Short of a few examples like providing a character's precise age, Barthes is unfortunately laconic in his explanation of this phenomenon. Regardless, from Barthes' writing we glean that any method of cataloguing details must encode the occurrences of familiarizing information which should be identifiable through cardinal directions, numbers, colors, and other material realities. To wit, a close-reading of Leo Tolstoy's realist fiction posits these sort of empirical details are significant because they contribute to a discourse that generates persons and worlds that feel authentic (Auyoung, 2015).

Meanwhile, fellow narratologist Gerard Genette opposes "narration" and "description" (Genette, 1976). Narration represents actions and events while description depicts objects and people (Genette, 1976). Description, according to Genette, "lingers over objects and ... seems to suspend the flow of time... successively [modulating] the representation of objects simultaneously juxtaposed in space" (Genette, 1976). This bifurcation parallels Viktor Shklovsky's dyad of "fabula" (story) and "syuzhet" (plot) where syuzhet serves as "a phenomenon of style" and allows the author with to augment a narrative core that can always be reduced to "Because A, Because B, etc." (Liveley, 2019).

Genette's argumentation mirrors Barthes in that specific cases of this conjectured "modulation" occurring with a syuzhet are scant. Lack of examples notwithstanding, Genette offers justification for contending that a novel's descriptive power lies in its capacity to vividly depict and fictional objects

(Genette, 1976). Other scholars have attempted to extend this notion, such as advancing a formal conception of a "descriptive detail," but again this type of analysis is plagued by generalities and a lack of reproducible steps for identification (Schor, 1984).

2.2 Noteworthy Digital Narratology

The most relevant computational studies for this effort operationalize Barthes' theory of salience. For instance, a fascinating study takes Barthes at his word and operationalizes cardinal functions to determine whether the **BERT** language model can effectively detect important sentences across a corpus of folktales (Devlin et al., 2019; Otake et al., 2020). This study considers a sentence to be "salient" if its omission from the story greatly reduces the narrative's coherence. Furthermore, a duo of prolific computational narratologists are producing excellent studies regarding tasks such as identifying suspense (Wilmot and Keller, 2020). In regards to salience, the pair augments the technique used by Otake et al. (2020) with an enhanced model: a question-and-answer mechanism plus an attentive layer responsible for handling the increased contextual demands of their corpus (Wilmot and Keller, 2021). They contribute an original learning pipeline that can identify salient sentences across a corpus of novel-length works. Additionally, Ted Underwood's work studying how to measure narrative time pairs excellently with Genette's narratology (Underwood, 2018). For my interests, capturing the "amount" of time elapsed in a given passage would be a valuable data point for an such an analysis. The methods to produce Underwood's duration values rely on a refined annotation scheme, but there are syntactic and morphological markers of time passing (e.g. "the next day") to capture.

3 Work to Date: Data Collection and Analysis

3.1 Corpus Formulation, Passage Sampling, Annotation

I'm sourcing novels from *Project Gutenberg* using the `c-w gutenberg` wrapper with a script that automatically grabs digital texts (Wolff, 2021; Gutenberg).[1] I am currently working with 28 novels. I wrote a sampling script that generates a random number of samples of a particular length from a corpus. Using this process, I drew 13 random samples of 800 characters from each novel in the corpus. This length was determined after a process of reading passages of varying lengths and judging whether or not there was proper context to glean the level of "detail" inhering the text.

From there, I used the sampling code to randomly draw a sample from this group (of 364) that prompted me to input a detail "rating" (scaled from 1.0 to 5.0) for the passage. To affix the scores, I wrote a script that randomly pulls a sample from a directory and prompts the user to apply a rating. The name of the work and author is not visible during this evaluation, but since I have read almost all the works in my corpus (some more than once), I am mostly aware of what book or author the sample originated from — especially when icons like Fitzwilliam Darcy grace me with their presence. While considering a text's score, I paid close attention to the amount of distinct objects and places that were mentioned and described. Passages that summarize an event or consist of choppy, back-and-forth dialogue receive lower scores than a Linnaean description of a sea-creature from Jules Verne. Meanwhile, meditations on elements of the natural world or, say, a fine-grained characterization of a lampshade receive higher scores. Interestingly, four of the randoms samples explicitly use the word "detail," seemingly holding conversation with Barthes and Genette. For example, a passage monologue in Fyodor Dostoyevsky *Crime and Punishment* muses about "the faces of clerks absorbed in petty details" (Dostoyevsky). Authors are clearly aware of the tension between unravelling a plot and properly enmeshing a reader into the character's milieu.

3.2 Exploring the Ratings

After reading and tagging each passage, I now have a (first-pass) rating for how the level of "detail" they contain. The distribution of the scores is roughly binomial, which follows a general sensibility: passages are either pretty detailed, or not that detailed. The median rating is 3.3 and the mean rating is 3.1, indicative of a left-skew (as in, there are slightly more resoundingly "not" detailed passages compared to "surefire" detailed passages). There are 197 passages with ratings greater than the mean and 167 with ratings lower than the mean. While the ratings I generated are by no means perfect, they resolve into a constructive first attempt to make

[1] All code mentioned from here on, along with an itemized corpus and the current feature-matrix, can be viewed at github.com/timschott/details.

sense of my group of passages. Based on each samples relation to the mean (above or below), I thus construe a partition of pseudo "labels": `detail` and `not_detail`. These labels are used in later sections as "groups" to compare during aggregations and graphing. Again, they are certainly not authoritative, but they represent the theoretical end-product of a study premised on ratings sourced from repeatable, objective heuristics. At this stage, the divisions are an simply an imperfect yet positive first step towards a formulation of how to identify a detail.

3.3 Log-Odds

The first such experiment I run with my random samples implements the log-odds ratio with an informative prior (Monroe et al., 2008). The goal of this routine is to determine which words best represent my two pseudo-camps — as used to aplomb in existing research(Monroe et al., 2008; Jurafsky et al., 2014). For every word in the corpus, I carry out a frequency-based analysis (both per group and overall) and output a score for each word that represents its allegiance with one of the two camps. For example, the 3 most distinctive words aligned with `detail` samples are {the, of, with} while the 3 most aligned with `not_detail` samples are {you, is, be}. Interesting differences manifest from this slice alone; the presence of the second-person "you" suggests an outsized frequency of exchanges between particular characters and therefore signals that the sample at hand contains sparse detail. To continue, the "top-25" for the `detail` group contains words for colors {black, white}, numbers {two, three}, prepositions {through, behind}, naturalistic elements {wind, air} and man-made settings {house, room}. Different categories of details (colors, numbers, naturalistic phenomena) organically emerge as we contemplate this output. For instance, associating of prepositions with detail passages comports with the preposition's central function of revealing the relationship between relationally-paired nouns. Looking forward, it will be interesting to rerun this analysis with a larger dataset and observe whether these outcomes persist.

3.4 Specificity, Parts-of-Speech

My next experiment quantifies the "specificity" found in each sample. At first, one would likely surmise that passages trafficking in details would use highly specific words. Nelson (2020) provides a computational method for evaluating this theory. I use the `SpaCy` software package to part-of-speech tag and lemmatize an inputted sample (Honnibal and Montani, 2017). Then, for each (adjudged) noun and verb used in the sample, the distance from the word's lemma to its broadest hypernym is calculated via the `nltk` Wordnet API (Toolkit, 2021; University). Per Nelson (2020), I then average these distance calculations for each random sample. To simplify this procedure, I use the "first" (typically the most-frequent) synset found in WordNet (e.g. for house, `house.n.01`). This method is successful at roughly an equal rate to much more sophisticated word sense disambiguation schemes (Raganato et al., 2017). In this routine, I also record part-of-speech tallies for classes of interest: noun, verb, adjective, adverb and adpositions (a superset of prepositions).

In step with the log-odds calculations, inspections of the specificity results are quite enlightening. All 10 of the most specific samples are members of the `detail` group; 9 out of the 10 least specific samples belong to the `not_detail` group. The rogue sample comes from Daniel Defoe's *Robinson Crusoe*, whose earnest, idiosyncratic prose — its noun usage, especially — presents one of the more challenging strains of writing to classify (Defoe). The most specific passage in the corpus comes from Tolstoy's *Anna Karenina*, when the author describes the inside of church with vivid, anaphoric diction: "...and the rugs, and the banners above in the choir, and the steps of the altar, and the old blackened books, and the cassocks and surplices..." (Tolstoy). A passage like this illustrates that the specificity metric is a helpful data point to consider when hunting for details. While readers obviously cannot recreate the query-based work used to produce this metric, they should be advised to heed the granularity of an author's nouns and verbs (e.g. whether a character errands to a "store" or the more specific "fishmonger").

To continue, `detail` samples are 14% more specific than `not_detail` samples. The part-of-speech statistics explain this result. The hypernym chain for any given verb is likely to be much smaller than that of a noun. A commonly used verb such as "enter" is the broadest member of its synset, while a commonly used noun like "house" is 9 levels down from its broadest member, "entity." As such, a passage with more nouns than verbs will likely boast a high specificity score. On average, `detail` samples contain 17% more nouns than their counterparts

and, likewise, 17% more adpositions. With this in mind, we can orient an analysis of the diverging subject matter and diction of these groups about a novel axis; the `detail` samples deliver information about things, places, and states of being while the `not_detail` passages narrate events and kinetic activities. Moving forward, I am excited to pursue related syntactic lines of inquiry. For instance, what kinds of nouns are most prevalent across the corpus? Are objects ubiquitous in instances of detail like Genette would have us believe (Genette, 1976)? Investigations into preposition usage possess potential for fertile results as well.

4 Next Steps

4.1 Corpus and Feature Expansion, Guideline Formalization

I want to commence the next phase of the project by rerunning my experiments with a larger corpus. It is straightforward to inject more data to the routines I have been running thanks to automated scripts; scaling the log-odds and specificity scores work to more samples is likewise trivial. As such, I will investigate closer to 100 novels (mostly canonical, still) and draw more samples from them as well. Now, I do not think it is feasible or that it would be particularly revelatory to annotate these thousands of samples with a rating. Reading every passage and attaching a score was an extremely helpful exercise for me to consider what does and does not comprise a detail in the confines of a sample. However, looking ahead, I want to focus on creating interesting features using the words found in the samples. I am going to conduct a more specific parts-of-speech analysis assessing noun and preposition usage. Distributed word representations may be of service for this portion of the project because of their capacity for capturing semantic relationships (Heuser, 2016). I also want to explore and capture semantic formulations stemming from Underwood's conception of literary time (Underwood, 2018). Lastly, I also want to incorporate a modified version of the salience recognition mechanism originating in Wilmot and Keller (2021). Ideally, I could reverse the scoring system and gain an understanding of which samples are the least enlightening (with respect to the fabula).

From there, I'll use my (growing) bank of metrics as the basis for formulating detail identification guidelines. Privacy scholars provide an excellent model for this activity when they fashion detailed rubrics for interrogating the compliance of terms-of-service agreements with respect to extant privacy law — a semantically challenging task not unlike mine (of Education, 2015). My goal is to draft a document of similar legibility and lucidity with fictional details at the core.

4.2 Evaluation Strategy

To evaluate the framework I furnish, I can tally the inter-annotator agreement, through the use of Cohen's Kappa or other statistic, of participants that read the same passage with my guidelines as a reference. The higher the scores for two people reading the same passages, the better the recommendations should generalize to a large-scale (McHugh, 2012). The criteria I proffer will of course be tweaked, iterated and remolded based on the helpful recommendations of other like-minded scholars interested in applying computational modes of study to literary phenomena.

5 Conclusion

To give Barthes (1989) one last due, let's ponder a spellbinding question which summarizes my research motivation: "If everything in narrative is significant, and if not, if insignificant stretches subsist in the narrative syntagm, what is ultimately, so to speak, the significance of this insignificance?"

Mass-scale automation of artistic endeavors is just around the corner, yet an innovative study only mentions the literary "detail" a single time (Goldfarb-Tarrant et al., 2020). Devoting time to the "significance of [the] insignificance," it would seem, is not in vogue. I believe the field's current emphasis on salience and summarization makes the task of taxonomizing and treasuring a (human-produced) novel's vast collection of details all the more poignant (Gupta, 2020; Liu, 2019).

My work will produce actionable insights regarding the implementation of details in fiction and provide numerous avenues for further study with the furnishing of a set of guidelines stating precisely how readers and digital systems alike can recognize and quantify details given a body of text. The goal of the project is not the production of a state-of-the-art model capable of recognizing what is and is not a detail, rather, to create the conditions in which constructing such a model is for the first time a real possibility.

References

Erich Auerbach. 2003. *Mimesis: the representation of reality in Western literature*. Princeton University Press.

Elaine Auyoung. 2015. Rethinking the Reality Effect. ISBN: 9780199978069.

Roland Barthes. 1989. *The Rustle of Language*. University of California Press.

Roland Barthes and Lionel Duisit. 1975. An Introduction to the Structural Analysis of Narrative. *New Literary History*, 6(2):237–272. Publisher: Johns Hopkins University Press.

Daniel Defoe. *The Life and Adventures of Ronbinson Crusoe*. Seeley, Service & Co. Limited, London.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Fyodor Dostoyevsky. *Crime and Punishment*.

U.S. Department of Education. 2015. Protecting Student Privacy While Using Online Educational Services: Model Terms of Service. *U.S. Department of Education*, page 9.

Gérard Genette. 1976. Boundaries of Narrative. *New Literary History*, 8(1):1–13. Publisher: Johns Hopkins University Press.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content Planning for Neural Story Generation with Aristotelian Rescoring. *arXiv:2009.09870 [cs]*. ArXiv: 2009.09870.

Shashank Gupta. 2020. On application of Bayesian parametric and non-parametric methods for user cohorting in product search. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 86–89, Seattle, WA, USA. Association for Computational Linguistics.

Project Gutenberg. Project Gutenberg.

Ryan Heuser. 2016. Word Vectors in the Eighteenth Century, Episode 1: Concepts.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Genevieve Liveley. 2019. Russian formalism. In *Narratology*. Oxford University Press, Oxford.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

David H. Miles. 1979. Reality and the Two Realisms: Mimesis in Auerbach, Lukács, and Handke. *Monatshefte*, 71(4):371–378. Publisher: University of Wisconsin Press.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Laura K. Nelson. 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 49(1):3–42. Publisher: SAGE Publications Inc.

Takaki Otake, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi, and Kentaro Inui. 2020. Modeling event salience in narratives via barthes' cardinal functions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1784–1794, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural Sequence Learning Models for Word Sense Disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Naomi Schor. 1984. Details and Realism: Le Curé de Tours. *Poetics Today*, 5(4):701–709. Publisher: [Duke University Press, Porter Institute for Poetics and Semiotics].

Leo Tolstoy. *Anna Karenina*.

Natural Language Toolkit. 2021. Natural Language Toolkit (NLTK). Original-date: 2009-09-07T10:53:58Z.

Ted Underwood. 2018. Why Literary Time is Measured in Minutes. *ELH*, 85(2):341–365.

Princeton University. Princeton University "About WordNet.".

David Wilmot and Frank Keller. 2020. Modelling Suspense in Short Stories as Uncertainty Reduction over Neural Representation. *arXiv:2004.14905 [cs]*. ArXiv: 2004.14905.

David Wilmot and Frank Keller. 2021. Memory and Knowledge Augmented Language Models for Inferring Salience in Long-Form Stories. *arXiv:2109.03754 [cs]*. ArXiv: 2109.03754.

Clemens Wolff. 2021. c-w-gutenberg. Original-date: 2021-03-29T04:32:36Z.