Quantifying Perceived Distance between Spatial Entities in Literary Text

Rosie Larson and Sandeep Soni
Quantitative Theory and Methods
Emory University
{rosie.larson,sandeep.soni}@emory.edu

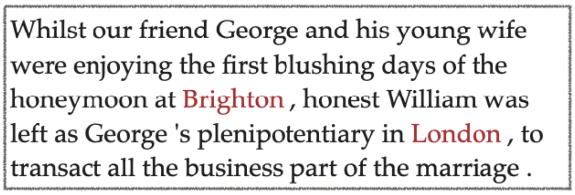
Abstract

Space is a literary device used by authors to structure their narrative. Authors make stylistic and narrative choices when describing, for example, the movement of characters between two locations as *a tiring car ride* or *a quick flight*. How do readers perceive the distance between spatial entities in narrative text? In this paper, we describe the task of estimating the perceived distance between masked spatial entities from a passage. We annotate passages selected from publicly available novels that are part of Project Gutenberg and contain two disconnected Geopolitical entities (GPEs). Our initial annotations suggest that though there is correlation of the perceived distance with the actual distance between the entities, there is also variance in the perceived distance, suggesting that the authors treat the spatial entities to modulate reader perceptions.

1 Introduction

Literary scholars have theorized the importance of geographical space in the construction of narrative meaning (e.g., Moretti, 1999; Piatti et al., 2009; Ryan et al., 2016). Both the "setting" (the backdrop of a story or a scene) and the movement in space of characters have been identified as critical narrative elements (Piper et al., 2021). Authors make deliberate stylistic choices in describing spatial entities in a narrative that not only helps relate the entities in the geographical space but also in the more subjective cultural space.

The theoretical interest in the spatial elements of a narrative has been coupled empirical interest in recent recent years. To enable computational analysis of spatial relationships at scale, methods have been proposed to identify and distinguish between spatial entities in literary texts such as natural locations (e.g., *river*), facilities (e.g., *building*), and geopolitical entities (e.g., *England*) (Bamman et al., 2019). Additionally, annotation schemes to encode the topological relations between spatial entities such as whether a spatial entity is disconnected from another entity or partially overlaps another entity have been proposed (e.g., Mani et al., 2008; Pustejovsky, 2017). Recent work has shown that large language models can be used to ground characters to locations and distinguish between a character being in, near, or moving towards a location (Soni et al., 2023).



Whilst our friend George and his young wife were enjoying the first blushing days of the honeymoon at Brighton , honest William was left as George 's plenipotentiary in London , to transact all the business part of the marriage .

Figure 1: Passage that mentions two spatial entities (highlighted) and asks the reader to estimate the distance between the two entities. During the annotation, the highlighted entities are masked out and replaced by `Loc1` and `Loc2` tokens. Example is taken from William Thackeray's novel `Vanity Fair`.

We build upon the past empirical computational research in this work by proposing the task of measuring the distance between a pair of spatial entities as perceived by the reader. Specifically, given a passage with two spatial entities (see Figure 1), we ask the reader to estimate the distance between the entities on a scale(see Table 1). Measurement of perceived distances can offer a lens into the stylistic choices employed by the author in constructing the narrative, and as a step in quantifying the movement in space. For example, a lack of cultural similarity between *New York* to *Ithaca* may lead the reader to estimate the distance between the two as significantly greater than the distance between *New York* to *Boston*, even though the geographical distances are roughly the same. Moreover, the portrayal of spatial entities in text may also mirror changes in the modes of transport, which could be seen in the perceived distance between any entities.

In this work, we describe our early progress on

Rating	Distance(in miles)	Example
1	0-1	NA
2	1-50	San Francisco and San Jose (40 mi.)
3	50-500	DC and NYC (200 mi.)
4	500-2,000	Miami and NYC (1100 mi.)
5	2,000-4,000	San Francisco and Miami (2600 mi.)
6	4,000+	L.A. and London (5,400 mi.)

Table 1: Proposed scale for annotating perceived distances. The scale was selected by running pilot annotations over small samples of data. Passages containing incorrectly marked spatial entities were given a “N/A” label

this task.

2 Data

We randomly selected passages from publicly available books present in Project Gutenberg.[1] We further filtered the passages according to the following criteria:

- The passage should contain at least two distinct entity mentions that are marked as proper noun Geopolitical entities (GPEs; e.g., *London*, *France*, etc) by BookNLP.[2]

- The passage has 25 tokens before and after the mention of the first and the second spatial entity, respectively.

3 Task

We randomly selected passages that intersected with the LitBank collection (Sims et al., 2019) and annotated 98 passages that matched the criteria mentioned before. Before annotating passages for perceived distance, we only retained passages which had a disconnected location pair. Disconnected location pairs were defined to be spatial entities that do not have any overlap. For example, *Paris* and *France* are considered to be a connected location pair whereas *Paris* and *London* are not.

Next, the annotators were shown a passage and asked to estimate the distance on a scale given in Table 1. Every passage was presented by replacing the true spatial entity mentions with `<LOC1>` and `<LOC2>` tags so that the annotators rely only on the linguistic clues seen in the passage. Additionally, all the passages were shuffled so that passages with the same location pair are only grouped by chance.

Location Pair	# Annotations	Mean Rating	Std
Paris/London	22	3.45	0.510
England/France	12	3.75	0.622
France/London	11	3.45	0.522
France/Germany	10	3.80	0.632
France/Italy	10	3.90	0.316
Germany/Italy	9	3.89	0.333
Brighton/London	9	3.11	0.601
Switzerland/Paris	8	3.63	0.744

Table 2: Distance ratings for 8 unique spatial entity pairs in the annotated passages

4 Initial Findings

Table 2 summarizes our initial findings. Of the 98 passages, 7 were marked "N/A" due to one or more location names referring to characters. The mean rating is correlated with geographical distance. The pearson correlation between the perceived distance estimates and the geographical distances between the location pairs is 0.96, suggesting that annotators were able to judge the scale of the distance from the linguistic context even though the entities were masked. *Brighton* and *London* are the closest locations geographically in our annotation dataset, and also yield the lowest mean rating of 3.11. The standard deviation of each location pair indicates that this task is non-trivial in that one location pair can yield different scale ratings based on the individual contexts of the passages in which it occurs.

A close inspection of another passage with the *Brighton-London* pair reveals contrast with the example in Figure 1.

> He felt he would go mad if he had to spend another night in London. Mildred recovered her good temper when she saw the streets of Brighton crowded with people making holiday , and they were both in high spirits as they drove out to Kemp Town. ...
>
> (William Somerset, *Of Human Bondage*)

While both passages contain the same location pair, the scale ratings differ because of the context provided. The passage above is rated 2 because the characters are assumed to have driven between the locations in a short time period. For the passage in Figure 1, the locations appear to be further apart because one location is a honeymoon destination, which is often far from home.

[1] https://www.gutenberg.org/

[2] https://github.com/booknlp/booknlp

Limitations

The preliminary findings from this paper, though encouraging, should be interpreted cautiously due the limitations of our setup. First and foremost is the small size of our dataset which forms the basis of our findings. In the future, we'll overcome this limitation in two ways: one, by expanding the annotation set; and, two, by building a predictive model for the perceived distance form the linguistic context.

Our focus on only GPEs for our initial analysis is limiting. GPEs represent only a subset of all locations mentioned in text, and it would certainly be beneficial to have a model trained on a variety of location types. In the context of our preliminary annotations, examining only frequently mentioned GPE pairs did not capture the full diversity of location pairs in literature. For example, no instances of the rating 1 and only one instance of the rating 2 were annotated. While many GPEs representing smaller cities, neighborhoods etc. and certainly many non-GPE locations would be within 50 miles of each other, narrowing our initial search to frequently mentioned GPE pairs skewed our results to larger scales. Going forward, it will be necessary to reevaluate our threshold for frequency and consider how this work could apply to non-GPEs.

Finally, the readers perception of distance could be influenced by factors like the historical and geographical context of the passages, which could pose issues for our annotations. Due to the restrictions on open access books offered by Project Gutenberg, much of the selected literature is over 100 years old and based in Europe. This presents issues in terms of the diversity of our annotation dataset, as well as how it might influence the reader's annotation. Knowing that the text is old based on the style could influence an annotator to think the scale of the locations mentioned is smaller. In other words, the reader may be aware of how distance is construed differently across historical time, and correct for it, leading to a a reduced result in our analysis.

Ethics Statement

Our work uses data from publicly available novels that are not under copyright anymore. However, historic data of this kind is prone to contain systematic biases; a majority of novels are written by White male authors in a Victorian setting with locations in the global North referenced much more frequently. Any findings should consider the historical context of the period that is analyzed. Our analysis is also restricted to English novels and so any findings about literary geography should be qualified to novels written in English.

References

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner. 2008. SpatialML: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Franco Moretti. 1999. *Atlas of the European novel: 1800-1900*. verso.

Barbara Piatti, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright. 2009. Mapping literature: Towards a geography of fiction. In *Cartography and art*, pages 1–16. Springer.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Pustejovsky. 2017. Iso-space: Annotating static and dynamic spatial information. *Handbook of linguistic annotation*, pages 989–1024.

Marie-Laure Ryan, Kenneth Foote, and Maoz Azaryahu. 2016. *Narrating space/spatializing narrative: Where narrative theory and geography meet*. The Ohio State University Press.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Sandeep Soni, Amanpreet Sihra, Elizabeth Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11723–11736, Toronto, Canada. Association for Computational Linguistics.