Methods of Analyzing Semantic Trajectories of Literary Works

Kirill Chertoganov

Supervisor: Vasilii Gromov

Researchers analyze semantic trajectories

However, there are some **limitations**:

- Variability of literary data: The diversity of genres and styles complicates the standardization of methods.
- Interpretation of results: The ability to adequately and objectively interpret the results of semantic trajectory analysis is critically important for the reliability of research findings.
- Quality of source data: Errors in the data directly affect the reliability of the results.

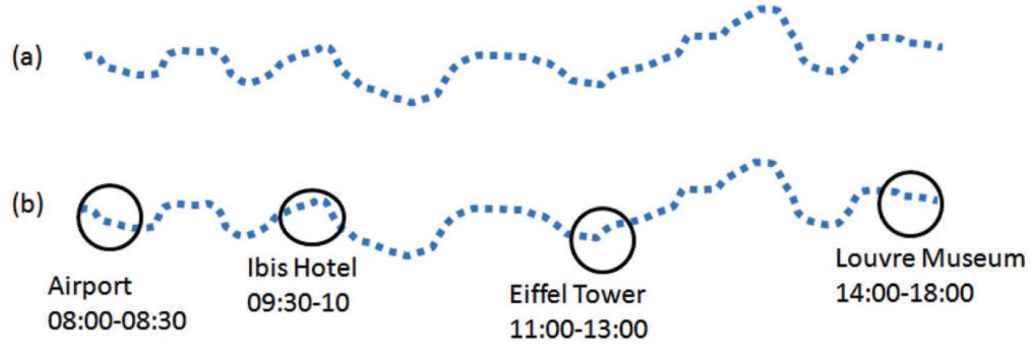


Figure 1. (a) Sample and (b) semantic trajectory

tid	stopid	stop_name	stop_gid	start_time	end_time	the_geom
2236	0	0_unknown	0	2004-08-04 18:10:18	2004-08-04 18:25:29	667015.36217975 7464208.41043153, ...
2234	0	1_unknown	0	2004-07-20 18:35:58	2004-07-20 19:06:29	679405.188285265 7469918.46315512, ...
2233	0	2_unknown	0	2004-07-20 15:39:33	2004-07-20 15:51:50	680479.071827787 7466287.37028003, ...
2232	0	3_unknown	0	2005-05-23 08:04:08	2005-05-23 10:06:25	673805.183177759 7455273.90615614, ...
1236	0	1_unknown	0	2005-02-21 17:20:42	2005-02-21 18:04:10	681063.492700548 7469816.57366467, ...
1235	0	1_unknown	0	2005-02-18 16:39:03	2005-02-18 17:08:24	681203.357858659 7469257.89224045, ...
1234	0	4_unknown	0	2005-02-17 15:33:33	2005-02-17 15:46:48	686860.9302822 7463298.01652945, ...
1233	0	1_unknown	0	2005-02-16 07:47:22	2005-02-16 08:11:14	680939.399430723 7469585.19999004, ...
1232	0	1_unknown	0	2005-02-15 16:45:57	2005-02-15 17:24:24	681037.05860043 7469842.71077073, ...
1231	0	1_unknown	0	2005-02-14 07:37:42	2005-02-14 08:24:26	679634.452558103 7471157.38290711, ...

Figure 2. Stops dataset

Typically, semantic trajectories modeling follows this framework

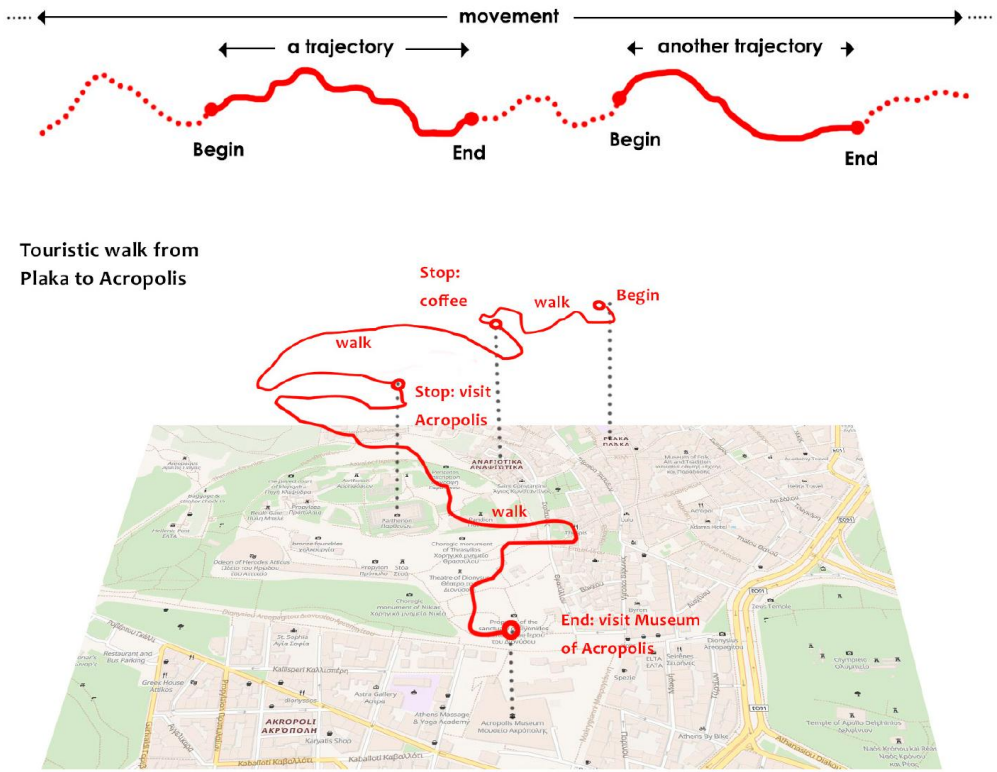


Figure 3. Example semantic trajectory depicting a tourist behaviour in Athens, starting from a simple trajectory, and resulting in a semantic trajectory

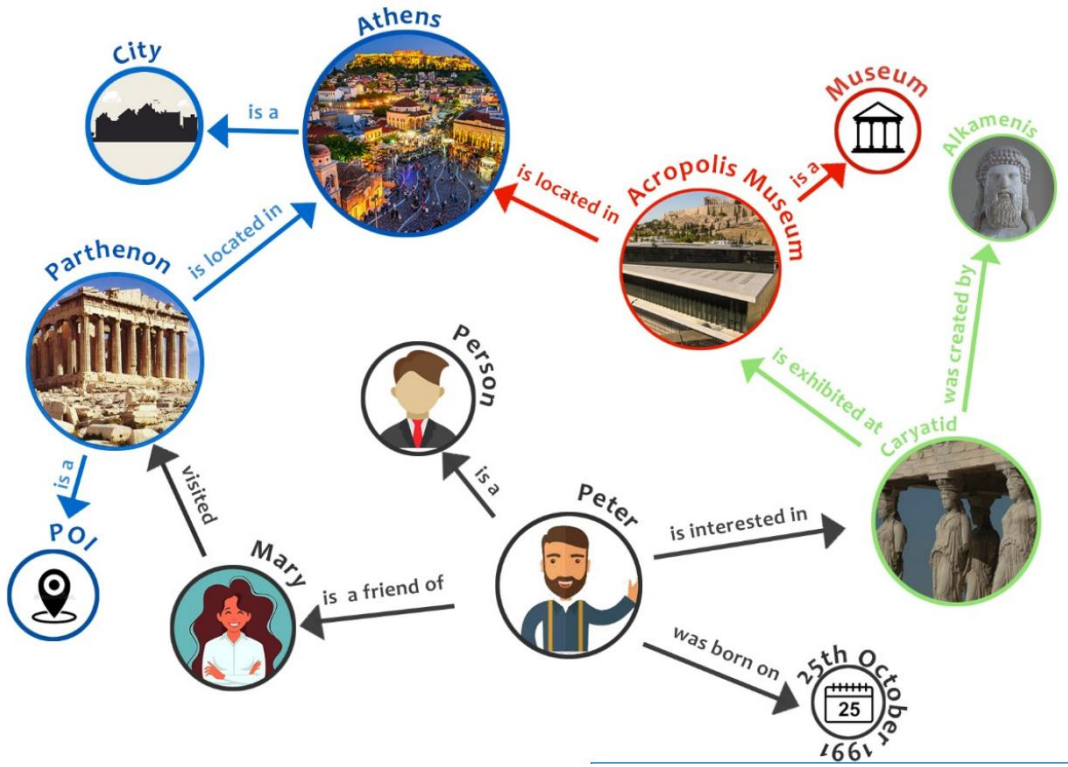


Figure 4. An example knowledge graph representing knowledge about the Acropolis and related entities, e.g., museums, visitors

Typically, semantic trajectories modeling follows this framework

Figure 3 outlines the proposed framework. At the base are data repositories, such as raw trajectory and spatial data, which are inputs. Preprocessing adds semantic information to these raw trajectories, generating a database for further analysis. The framework's core includes data preprocessing and mining modules, while the top layer focuses on visual data analysis, offering a graphical interface for users to preprocess, mine, and visualize trajectory data in map form. Preprocessing includes four key steps: cleaning trajectories, adding semantics, generalization, and formatting for data mining.

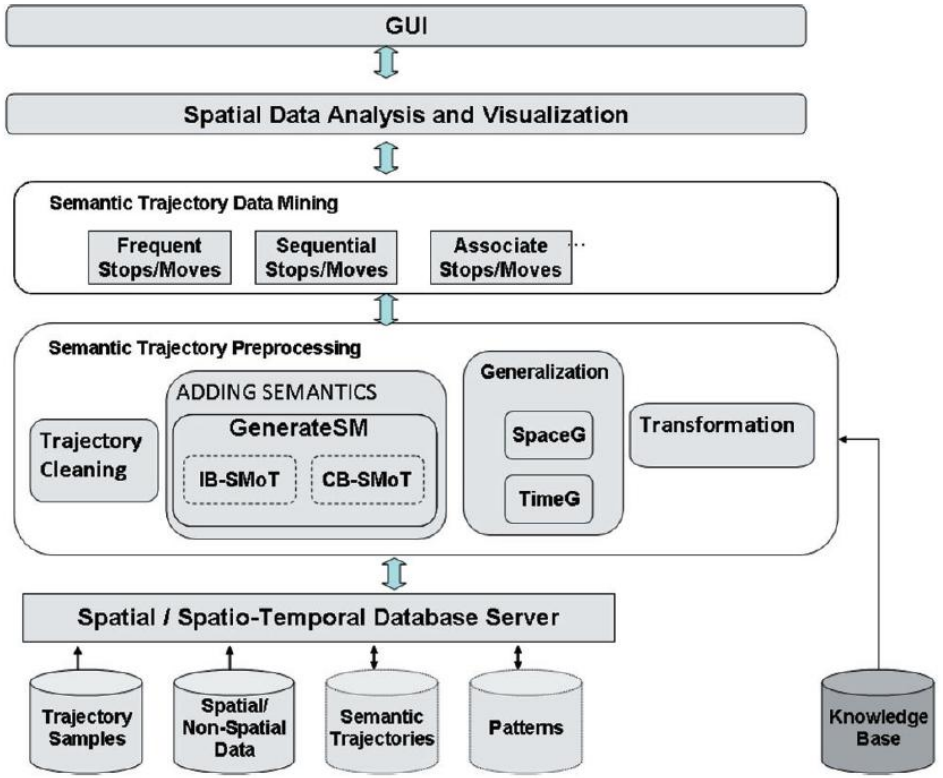


Figure 5. General software architecture

Research Group "Catch the Bot" Dataset

How to identify semantic features in literary works using advanced models?

This task involves analyzing thematic variations and context-dependent meanings.

We use a dataset from global literary repositories, applying state-of-the-art semantic analysis techniques.

Main question: can we find the best model to detect semantic trajectories in texts?

Hypothesis: The T5 (Text-to-Text Transfer Transformer) model, which unifies NLP tasks into a text-to-text format, provides the highest accuracy for detecting semantic patterns.

World literature corpus extracted from global database

Figure 6. Perseus Digital Library provides access to texts of ancient works

What’s new (mT5 + "Catch the Bot" data)

mT5 covers up to 101 languages, utilizes extensive datasets like Common Crawl, supports 300M to 13B parameters, and has the potential for improved performance in low-resource languages with further refinement.

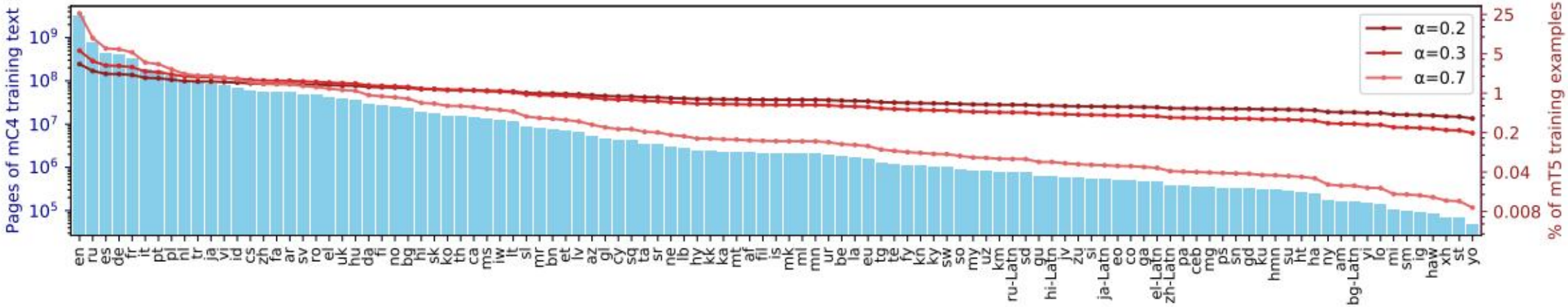


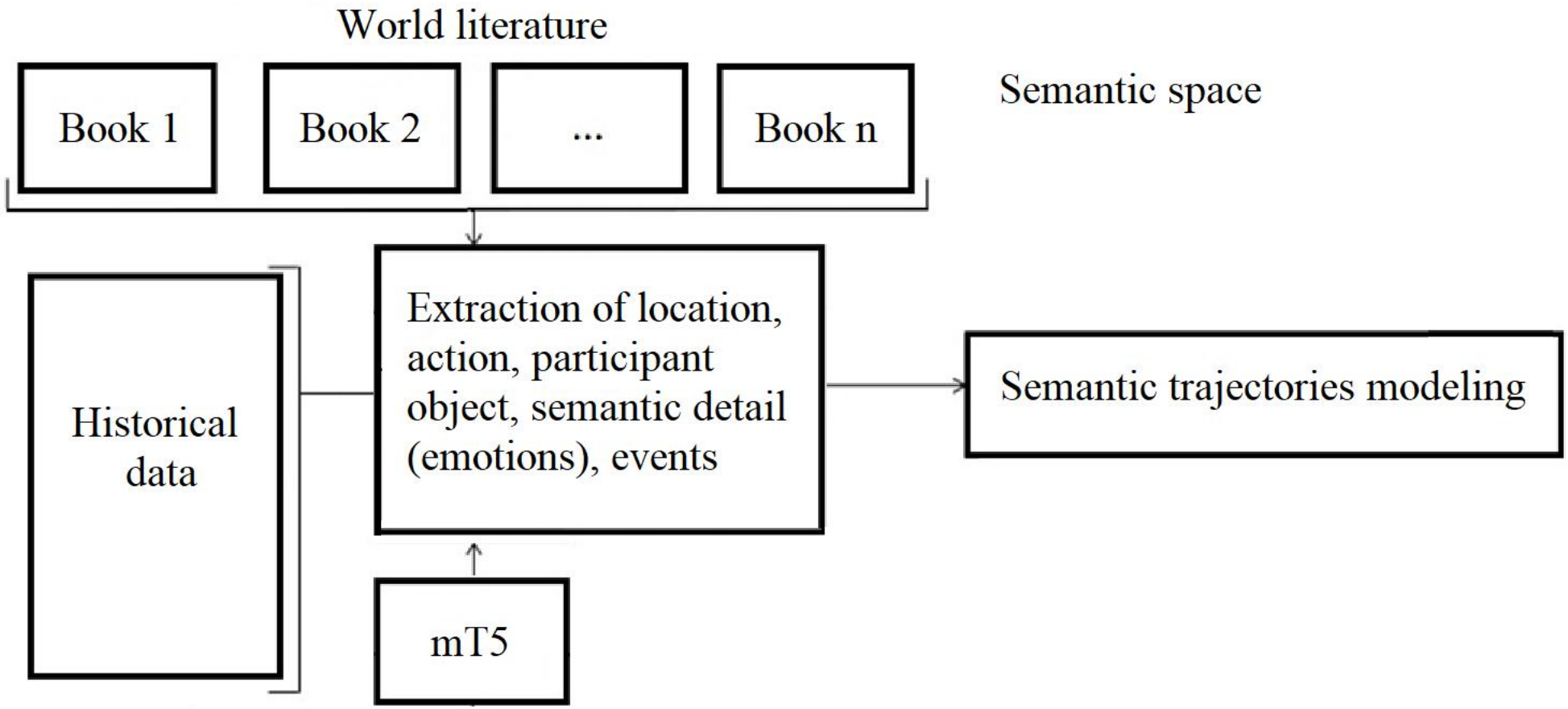
Figure 7. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

Metrics

Average F1: It is important for evaluating how well the model can detect and correctly interpret various meanings and themes in the text. This metric helps assess how accurately the model captures key **thematic elements** and **contextual variations**, even if they don't exactly match expectations.
EM (Exact Match): It is relevant when an exact match between the model's output and the original texts is required. This is especially crucial for tasks that need precise **interpretation of meanings** and detailed analysis, which is essential in literary studies.
Both metrics help understand how well the model handles tasks related to text analysis, where variability and accuracy of interpretation are key factors.

Task	Dataset	Model	Metric Name	Metric Value
Reading Comprehension	MuSeRC	MT5 Large	Average F1	0,844
Common Sense Reasoning	RuCoS	MT5 Large	Average F1	0,57
Common Sense Reasoning	RuCoS	MT5 Large	EM	0,562

Semantic trajectories modeling scheme



Developed Methods for Analyzing Semantic Trajectories of Literary Works

1 Established a model for literary analysis that identifies thematic trajectories and segments texts, seamlessly integrated into analytical software.

2 Created a new algorithmic framework capable of detecting semantic features in literary texts.

3 Successfully combined textual and contextual data to train the algorithm, enabling it to recognize and interpret complex semantic patterns effectively.

Prospects

1 The analysis of semantic trajectories in literature allows scientists to develop new approaches and tools for studying linguistic and cultural changes in texts, thereby expanding the boundaries of linguistics and literary studies. This contributes to a better understanding of both historical and contemporary literary processes and trends.

2 Using these methods can enrich cultural perception and understanding of literary movements and changes in language.

3 Semantic trajectory analysis can reveal communication patterns in global conflicts and diplomacy, aiding the UN's goals of enhancing security and stability by providing insights into international relations and geopolitical dynamics.