University *of Ljubljana*
Faculty *of Computer and Information Science*

Natural language processing course Latex Template

Matej Bevec, Nejc Hirci, and Jakob Petek

Abstract

Keywords
event salience, narrative understanding, sentiment analysis

Advisors: Slavko Žitnik

Introduction

Narrative understanding is a critical task in natural language processing that aims to analyze and comprehend the essential elements of a story, including characters, events, settings, and their semantic and temporal relationships [1]. Several approaches have been proposed for analyzing narrative progression, including the analysis of the progression of narrative features over time. There is theoretical background that suggests fluctuations in sentiment over time are a good proxy for fluctuations in plot movement. For example, a large body of work by the acclaimed novelist Kurt Vonnegut, describes stories precisely in these terms [2].

We focus specifically on *event salience* as a crucial narrative feature. Event salience refers to the significance of an event or passage and its importance to the plot. Since Slovene has been the subject of little attention in the field of natural language processing, we put emphasis on investigating how the approach at hand performs on this language.

We apply a cross-lingual salience detection framework Barthes' definition of event salience [3] to analyze narrative progression of a number of Slovene and English novels.We aim to evaluate the capacity of this approach to analyze long-form narratives by matching the obtained salience arcs to ground-truth annotated narrative events and classify them in terms of theory-backed story archetypes. Moreover, we evaluate the model's performance on English and Slovene languages

1. Related Work

Narrative understanding research [1] has focused on various aspects of story comprehension, including named entity recognition (and similarly, agent detection), relation detection, event detection, time detection, and setting detection. Particularly relevant to our work are the developments on narrative event extraction from temporal salience patterns. This section briefly describes these works and outlines NLP tools and resources that are available in Slovene.

In the paper Modeling Event Salience in Narratives via Barthes' Cardinal Functions [3] researchers proposed a model that estimates salience without any annotations by adopting Barthes' definition of event salience. The model estimates a sentence's salience by computing the amount of coherence loss when events in a sentence are removed from the narrative. The coherence loss is computed by using a pre-trained language model to estimate the generation probability of a sentence given a narrative. This approach can be applied to different languages by fine-tuning the pre-trained model on a chosen domain or using an appropriate embedding model. Building on this approach, we propose using event salience calculation to estimate the narrative structure in Slovenian and English novels. By identifying the salient events in the text and mapping them onto a graph, we can estimate the structure of the narrative and provide insights into the novel's plot.

It is worth mentioning that the temporal event salience method was significantly extended in the paper Memory and Knowledge Augmented Language Models for Inferring Salience in Long-Form Stories [4], where the text is augmented by a knowledge base (i.e. an ontology) to infer long-term semantic relationships that may sometimes be needed to fully comprehend an event's salience. This is however out of the scope of this paper, because it would require training a Slovene knowledge base based on Retrieval-augmented generation architecture and then further extend and fine-tune it with a memory mechanism.

To date, there has been significant progress in the development of natural language processing (NLP) tools for Slovenian language. One such tool is CLASSLA [5], which enables var-

ious levels of processing such as tokenization, part-of-speech tagging, lemmatization, dependency parsing, and named entity recognition. Importantly the previously mentioned approach based on Barthes' defintion of event salience requires a language model embedding to calculate narrative coherence loss per sentence, which is why we use the cross-lingual pretrained language model called CroSloEngual BERT 1.1 [6] to be able to compare the quality of our approach on both Slovene and English texts. This language model was trained on large monolingual corpora for English, Slovene, Croatian, but with bigger focus on Slovene and Croatian, in order to prevent English from overwhelming the other language models. We selected popular Slovene and English novels from the repository ELTeC [7], which includes a 100 Slovenian and English novels, with clearly established narrative structures, which we can manually annotate for evaluation.

Another important aspect with which we extend the event salience extraction method is by using named entity recognition and relation extraction, in order to observe event salience in a narrative on a per character basis. This can be easily achieved by utilizing available CLASSLA framework, which relies on the previously mentioned CroSloEngual BERT model.

In summary, this article presents a novel approach to event salience detection in Slovenian language and proposes its application to estimating the narrative structure in Slovenian and English novels. Our approach builds on previous work in this area and leverages the availability of NLP tools and resources for Slovenian language.

References

[1] Andrew Piper, Richard Jean So, and David Bamman. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, 2021.

[2] Matthew L Jockers. Revealing sentiment and plot arcs with the syuzhet package. 2015.

[3] Takaki Otake, Sho Yokoi, Naoya Inoue, Ryo Takahashi, Tatsuki Kuribayashi, and Kentaro Inui. Modeling event salience in narratives via barthes' cardinal functions. *arXiv preprint arXiv:2011.01785*, 2020.

[4] David Wilmot and Frank Keller. Memory and knowledge augmented language models for inferring salience in long-form stories. *arXiv preprint arXiv:2109.03754*, 2021.

[5] Nikola Ljubešić, Luka Terčon, and Jaka Čibej. The CLASSLA-stanza model for morphosyntactic annotation of standard slovenian 2.0, 2023. Slovenian language resource repository CLARIN.SI.

[6] Matej Ulčar and Marko Robnik-Šikonja. Crosloengual bert 1.1. 2020.

[7] Christof Schöch, Tomaž Erjavec, Roxana Patras, and Diana Santos. Creating the european literary text collection (eltec): Challenges and perspectives. *Modern languages open*, 2021.