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

Literacy situation models knowledge base creation

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 various 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' definition of event salience requires a pre-trained large language model in order to calculate a per-sentence narrative coherence score. Initially we used the well known Generative Pre-trained Transformer (GPT-2) model [6] for the English texts and a variant of the model trained on the Slovene texts.

We selected a subset of the well known Grimm's fairytales in English and Slovene language, which have very simple and clear narrative structures, which we will manually annotate by following the established guidelines of the Russian writer Vladimir Propp, which described different elements of folktales in his work *Morphology of the Folktale* [7], which also formed the basis for the previously mentioned unsupervised approach for event salience estimation.

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.

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.

2. Dataset

In the scope of our project, we apply our analysis to set of original brothers' Grimm's folktales. The choice of using folktales, instead of other narrative works, is motivated by the following considerations:

- Many folktales are sufficiently short to both to fall within the used language model's token limit and to be annotated manually without outsourcing the work.
- Their narrative structure tends to be simple and explicit, without special cases, such as "story withing a story" or unreliable narration.
- Many Grimms' stories in particular are translated to both English and Slovene, allowing for comparison.

We sourced 63 English translation of the original Grimms' fairytales from [8] — a dataset extracted from the Gutenberg project's free Grimms' fairytales e-book [9]. The Slovenian translations were sourced from [10] and include only 26 stories.

Because the Slovenian dataset does not contain all titles in the English dataset and vice versa, we took the intersection of both as our dataset. The resulting 15 story pairs are listed in Table 1.

After analyzing the story pairs, we noticed that there is a significant mismatch in the length of the English and the Slovenian translation for many of the considered stories (see Table 1). This is because the Slovenian collection appears to contain many shortened reinterpretations of the original folktales, while the English texts are all the originals (i.e. their translations).

In this iteration of the report, we choose to analyize some of the well-matched pairs. However, our agenda is to collect our own Slovenian dataset of Grimm's fairytales from physical prints via scanning and OCR.

2.1 Annotations

We expect our salience analysis model to produce a continuous salience estimate for each sentence in a story. To evaluate this output, both qualitatively and quantitatively, we also manually annotate each sentence.

We base our annotations around the theory of Vladimir Propp, outlined in his work *Morphology of the Folk Tales*. Propp proposes that narrative structure of most folk tales can be broken down into 31 generic narrative units called *narratemes*. These can be seen as common categories of salient events, such as "hero leaves on mission (departure)" or "villain is defeated (victory)". We much simplify this framework and consider any event as salient, if it roughly corresponds to any of the 32 narratemes. We annotate the text per sentence with binary labels. If a sentence contains or is part of a salient event, defined in above terms, it is labeled with with 1, otherwise, it is labelled with 0.

3. Methods

In our work we analysed the the structure of folk tales by following the approach described by Otake et. al. (2020) to estimate a per-sentence event salience. Preliminary results of our work can be seen Section 4. Further we will use this approach along with named entity recognition and corefrence resolution pipelines in order to extract a per-character narratives in each story.

3.1 Event salience

In their work Otake et al. (2020) proposed a unsupervised approach to measuring per-sentence event salience by levering the learned language understanding of large pre-trained language models. Following from Barthes' definition of events, which are defined as logically essential elements to the narrative action, the researchers proposed a simplification of event extraction by focusing only on the sentence's salience in a narrative. For the computation of sentence's salience score we remove the sentence from the narrative and compute the coherence loss compared to the narrative with the included target sentence.

Their method proposed computing the generation probability of a narrative and regarding it as the narrative's coherence score. This comes down to a simple product of word generation probabilities, which are influenced by the number of words in the narrative. The estimated coherence score is then estimated as the average log-likelihood of all tokens

Table 1. The considered Slovenian and English translations of Grimms' stories and their corresponding lengths (in words).

English title	Length	Slovenian title	Length
The Golden Bird	2540	Zlata ptica	2470
Jorinda and Jorindel	1146	Jorinda in Joringel	748
The Straw, the Coal and the Bean	506	Slamica, ogeljček in fižolček	347
The Frog Prince	1202	Žabji princ	1003
Hansel and Gretel	2952	Janko in Metka	557
The Mouse the Bird and the Sausage	707	Mišek, ptička in klobasa	614
Little Red Cap (Little Red Riding Hood)	1118	Rdeča kapica	465
Rumpelstiltskin	1151	Špicparkeljc	996
Ashputtel	2608	Pepelka	1750
The Wolf and the Seven Little Kids	1110	Volk in sedem kozličkov	950
The Fox and the Cat	294	Lisica in mačka	101
The Golden Goose	1593	Zlata gos	953
The Seven Ravens	1014	Sedem krokarjev	1321
The Story of the Youth Who Went Forth to Learn What Fear Was	3815	O mladeniču, ki bi rad strah poznal	1392

following the target sentence, which was removed from the text. In our initial implementation we used the GPT-2 models trained separately for Slovene and English, we will be however looking into extending our approach with other generative models such as T5 model and perhaps other established text generation metrics, such as BERTScore or DiscoScore.

It was important to address the limitation of input length of any selected pre-trained language model, which was often too small to accept the full selected texts. The proposed solution by Otake et al. (2020) was to simply include the largest possible input for the context before and after the selected target sentence. This can prove to be problematic for longer texts, because it goes against the assumptions of the provided approach. We propose an alternative solution to better address this issue by using a sliding window on each target sentence, which computes the average over all the possible contexts before and after the target sentence, given the model's input limit. We describe our results in Section 4. Furthermore we are thinking of exploring other ways including larger contexts in the selected generaiton models.

3.2 Extracting character arcs

We will we further exploring the previously described method of event salience estimation on a per-character basis in each story by using the provided pipelines present in the CLASSLA framework.

3.3 Visualization and qualitative analysis

4. Results

4.1 Event salience over time

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] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[7] Vladimir Iakovlevich Propp. *Morphology of the Folktale*, volume 9. University of Texas Press, 1968.

[8] Grimms' Fairy Tales.

[9] Jacob Grimm and Wilhelm Grimm. *Grimms' Fairy Tales*. April 2001.

[10] Index of /text-semantics/data/grimmove-pravljice/.