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

Narrative salience in short-form fiction

Matej Bevec, Nejc Hirci, and Jakob Petek

Abstract

Keywords
event salience, narrative understanding, narrative structure 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 structure, 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 based on Barthes' definition of event salience [3] powered by three different modern language models to analyze narrative progression of a number of Slovene and English translation of brothers Grimms' folktales. We aim to evaluate the capacity of this approach to analyze short-form fictional narratives by matching the obtained salience arcs to ground-truth annotated narrative events and classify them in terms of theory-backed story archetypes, based on the work of literary theorist Vladimir Propp. Moreover, we evaluate the model's performance on English and Slovene languages. To enable the latter, we collect our own novel Slovene language Grimms' folktales dataset.

The following sections first describe the acquired folktale dataset and our methodology. We then present and discuss the obtained results.

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. We evaluate three different pairs of language models: Generative Pre-trained Transformer (GPT-2) model [6], BERT and T5. For all three we source the English variant to use on the English texts and a variant trained on a Slovene corpus to used on the Slovene translations (see Section 3).

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] (see Section 2).

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 originally sourced from [10] and included only 26 stories. However, 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. 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). This presented a problem, as we require approximately direct (sentence-to-sentence) translations for our comparison with the English titles.

This prompted us to create our own Slovene dataset. A subset of English titles, consisting of 15 stories, was scanned from a physical copy of the 1980 edition of the *Grimmove pravljice* book [**?**] using optical character recognition (OCR).

After clean-up, the dataset encompasses 15 story pairs listed in Table 1.

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. Due to resource constraints, we only label 8 of the analysed stories (see Section 4).

To aid our qualitative analysis, we also descriptively label the events which we subjectively deem most important and compared them with the global maxima within the approximated salience arcs.

3. Methods

In our work we analyse the the structure of folk tales by following the approach described by Otake et. al. (2020) to estimate a per-sentence event salience. The results of our work can be seen in Section 4

3.1 Event salience

In their work Otake et al. (2020) proposed a unsupervised approach for measuring per-sentence event salience by leveraging 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 coher-

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

English title	Length	Slovenian title	Length
The Old Man and his Grandson	252	Ded in vnuk	240
The Mouse, the Bird and the Sausage	707	Mišek, ptička in klobasa	660
Doctor Knowall	780	Doktor vseznalec	673
Old Sultan	875	Stari sultan	739
The Twelve Huntsmen	1080	Dvanajst lovcev	959
The Wolf and the Seven Little Kids	1110	Volk in sedem kozic	907
Little Red Cap (Little Riding Hood)	1119	Rdeča kapica	995
Rumpelstiltskin	1151	Špicparkeljc	951
Clever Elsie	1348	Brihtna Urška	1154
The Dog and the Sparrow	1350	Pes in vrabec	998
The Travelling Musicians	1368	Bremenski mestni godci	1088
King Grisly-Beard	1496	Kralj Drozd	1418
The Golden Goose	1593	Zlata goska	1372
The Four Clever Brothers	1710	Štirje izučeni bratje	1351
Hansel and Gretel	2952	Janko in Metka	2484
The Story of the Youth Who Went Forth to Learn What Fear Was	3815	Zgodba o človeku, ki je šel po svetu, da bi strah spoznal	3284

ence 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 following the target sentence, which was removed from the text. The sentiment of a given sentence is then defined as the difference between the estimated coherence in the case where the sentence is present in the query and when it is not. In our implementation we use three different langage model architecture pre-trained separately for Slovene and English, which we describe below.

It was important to address the limitation of input length of any selected pre-trained language model, which is 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 generation models.

3.2 Language models

We evaluate three state-of-the art pre-trained architecture in the role of the underlying language model in the framework described above. These are GPT-2, BERT (in two configurations) and T5. For GPT and T5, we use separately pre-trained English and Slovene models. For BERT, we use a crosslingual variant.

- **BERT.** Bidirectional Encoder Representations from Transformers is a family of pre-trained general purpose masked language models introduced in 2018. The model is tasked to predict the story after the evaluated sentence and coherence is estimated as log-likelihood loss against the ground truth (i.e. actual rest of the story). Embeddia's *crosloengual-bert* crosslingual model is used.
- **GPT-2.** General Pretrained Transformer 2 is pre-trained general purpose transformer language model, introduced in 2019. Coherence is estimated in the same manner as above. For English, we use the **base** mode. For Slovene, we use CJVT's *gpt-slo-base* variant.
- **T5.** (Text-To-Text Transfer Transformer is a multi-task language models trained by re-framing common NLP tasks to a text-to-text (seq2seq) format. The *summarize* instruction is used to generate text before coherence is estimated in the same manner as above. For English, we use the t5-small variant. For Slovene, we use CJVT's *t5-sl-small* variant.
- **NSP BERT.** We additionally evaluate BERT in a next sentence prediction (NSP) configuration, where the model is tasked to predict whether a pair of sentences represents subsequent sentences or not. Here we estimate salience by comparing how accurate the model is in classifying whether a given sentence has been removed or not.

3.3 Visualization

Due to the nature of our study, a large part of our analysis is qualitative and reliant on visualization. We represent approximated salience for a given story and a given model with line plot (in black). The x-axis represents time in sentences, while the y-axis denotes estimated salience for that sentence. Since the estimates are somewhat noise, we attempt to increase visual clarity by also visualizing the same data filtered through a frequency domain low-pass filter (in blue). The estimates are overlayed by the ground truth annotations. Red X marks represent sentences annotated with 0 (non-salient), while green dots represent sentences annotated with 1 (salient).

We describe the results based on these visualizations in Section 4.

3.4 Evaluating estimate accuracy

The second part of our analysis is quantitative. Here we devise a simple accuracy measure to evaluate obtained estimates against our ground-truth annotations. The simplest approach would be to determine, per-story or globally, a *constant threshold*. If the salience estimate falls above it, the sentence in question is considered salient and vice-versa. Then, these tresholded values can be compared to ground-truth with simple *classification accuracy*. However, it can be observed that while the estimates vary locally, from sentence to sentence, there is also present a low-frequency component which may be referred to as baseline wander in signal processing terms. This may unfairly reduce accuracy if a static threshold is used. Therefore, we choose to use a dynamic threshold — simply the low-pass-filtered salience curve, as described above (visualized in blue).

4. Results

In our preliminary results we observed the extracted event salience for each sentence in 3 manually labeled narratives and compared our results between the sliding window and maximum context approach on both English and Slovene versions.

In this section we describe our results obtained from the 6 evaluated models on the 4 annotated story pairs.

4.1 Estimate accuracy

Accuracy scores in Table 2 range from around 0.3 to 0.7, with a means for all models slighly above or below 0.5, indicating the approach performs rather poorly at least with respect to our metric. Since the ratio of annotated-salient sentences does not skew far from a one-to-one ratio, this indicates a barely better-than-random performance. However, there are noticable differences between the evaluated models.

Although there is significant variance in performance between different texts, we can deduce the following. The GPT-2 model performs the best, with consistent scores above 0.5 and a mean of 0.55 and 0.6 for English and Slovene respectively. Second is the T5, while surprisingly, BERT is the worst, often performing worse-than average. The windowed approach does not significantly alter accuracy. It increases it slighly for GPT and decreases it slighly for BERT, but this varies significantly between texts. Finally, both GPT-2 and T5 perform notably better in Slovene, while the crosslingual BERT performs similarly or worse. This may indicate that a purpose-trained model is more effective than a crosslingual one.

4.2 Event salience over time

Figures 1 and 2 depict the salience plots for the Little Red Hood story and its Slovene translation. The visualization largely confirms what the quantitative results indicate. Interestingly, the GPT-2 model and, to some extent, the T5 model, reveal a story arc with a salience peak near the end. The BERT models do not. This is a common across most evaluated stories. According to the former two models, Grimms' folktales tend to build up to a narrative resolution at the very end.

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.

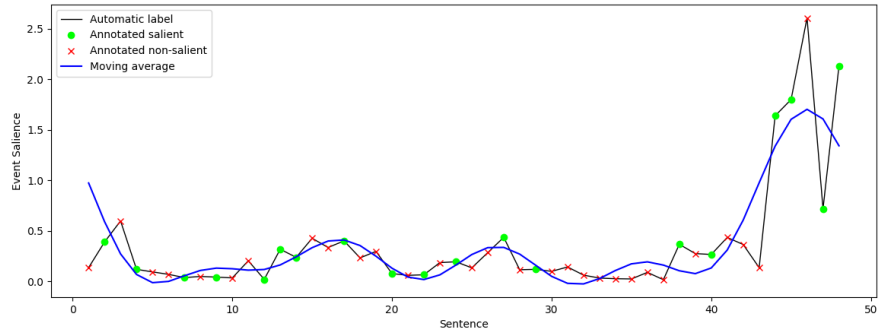
[8] Grimms' Fairy Tales.

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

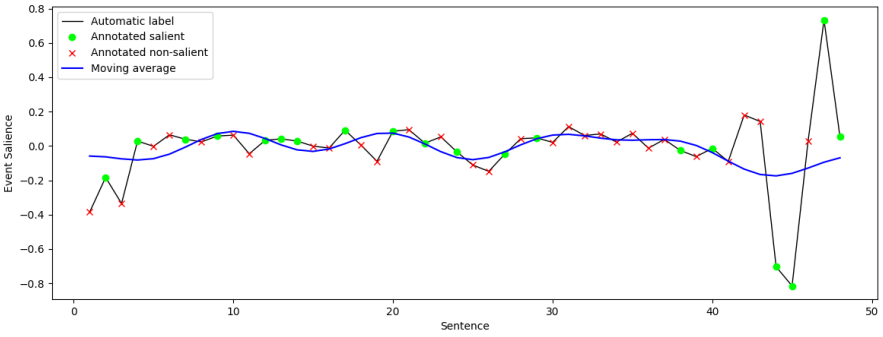
[10] Index of /text-semantics/data/grimmove-pravljice/.

Text	GPT-2	GPT-2 (window)	T5	BERT	BERT (window)	NSP BERT
Little Red Riding Hood	0.5	0.54	0.54	0.58	0.5	0.44
The Wolf and the Seven Little Kids	0.59	0.67	0.36	0.39	0.47	0.49
The Mouse, the Bird and the Sausage	0.54	0.42	0.38	0.54	0.46	0.46
Rumpelstiltskin	0.55	0.55	0.48	0.43	0.5	0.48
Mean (English)	0.55	0.55	0.48	0.49	0.48	0.47
Rdeča kapica	0.60	0.70	0.62	0.41	0.35	0.41
Volk in sedem kozic	0.69	0.66	0.57	0.44	0.42	0.44
Mišek, ptička in klobasa	0.58	0.62	0.48	0.52	0.42	0.48
Špicparkeljc	0.52	0.5	0.42	0.42	0.5	0.5
Mean (Slovene)	0.60	0.62	0.52	0.45	0.43	0.46

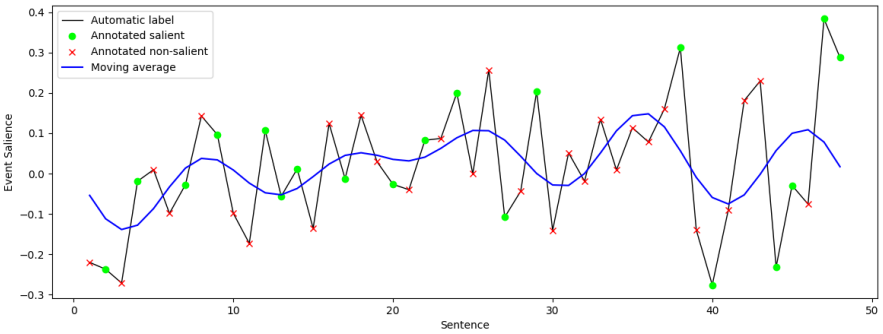
Table 2. Salience estimate accuracy for the annotated English and Slovene folktales.



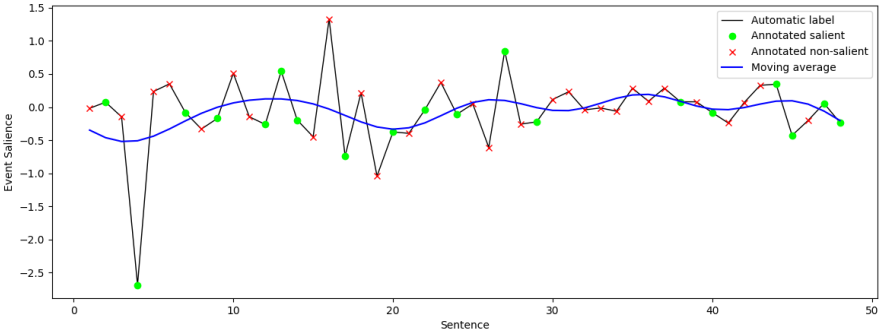
(a) GPT-2



(b) T5

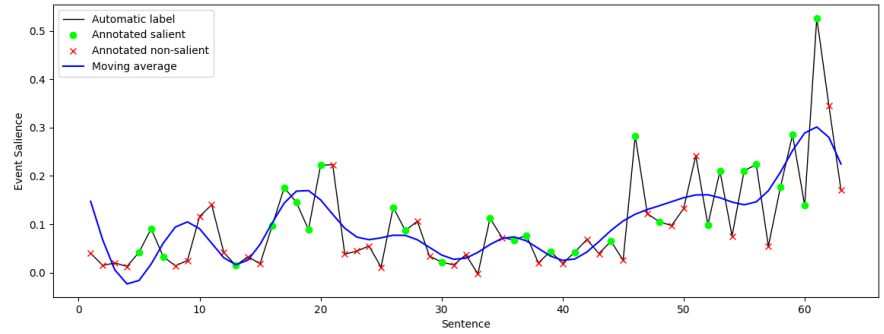


(c) BERT

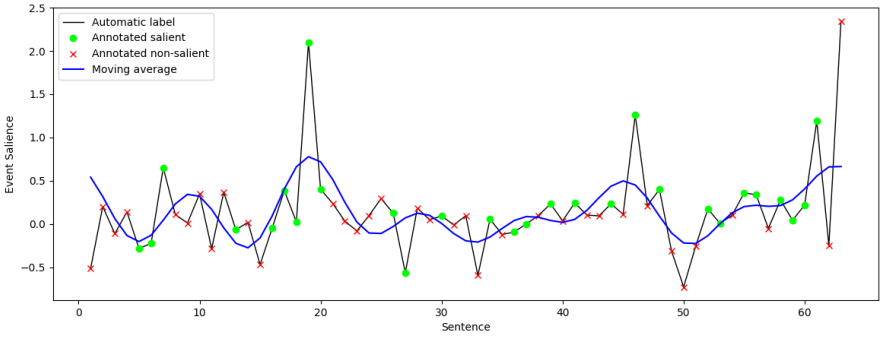


(d) NSP BERT

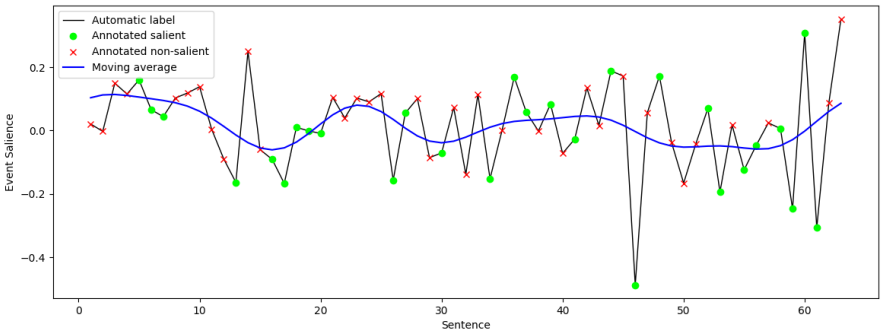
Figure 1. Estimated salience over time for the *Little Red Cap* story.



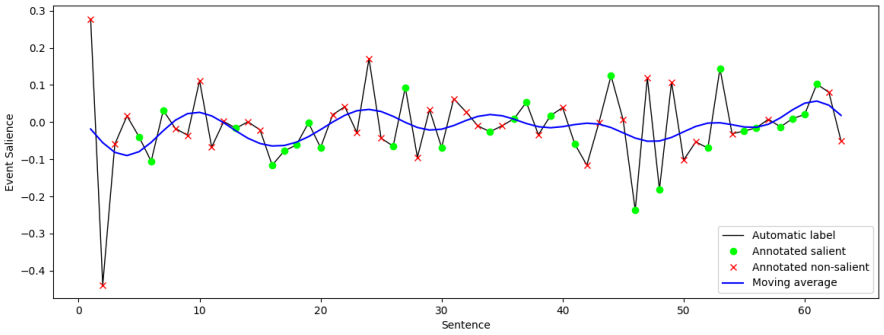
(a) GPT-2



(b) T5



(c) BERT



(d) NSP BERT

Figure 2. Estimated salience over time for the *Rdeča kapica* story.