What's New?
Identifying the Unfolding of New Events in a Narrative

Seyed Mahed Mousavi[*], **Shohei Tanaka**[†,‡], **Gabriel Roccabruna**[*],
Koichiro Yoshino[†,‡], **Satoshi Nakamura**[‡], **Giuseppe Riccardi**[*]
[†]Guardian Robot Project, RIKEN, Japan
[‡]Nara Institute of Science and Technology, Japan
[*]Signals and Interactive Systems Lab, University of Trento, Italy
mahed.mousavi@unitn.it,giuseppe.riccardi@unitn.it

Abstract

Narratives include a rich source of events unfolding over time and context. Automatic understanding of these events provides a summarised comprehension of the narrative for further computation (such as reasoning). In this paper, we study the Information Status (IS) of the events and propose a novel challenging task: the automatic identification of *new* events in a narrative. We define an event as a triplet of subject, predicate, and object. The event is categorized as new with respect to the discourse context and whether it can be inferred through commonsense reasoning. We annotated a publicly available corpus of narratives with the new events at sentence level using human annotators. We present the annotation protocol and study the quality of the annotation and the difficulty of the task. We publish the annotated dataset, annotation materials, and machine learning baseline models for the task of new event extraction for narrative understanding.

1 Introduction

The task of narrative understanding is a challenging topic of research and has been studied in numerous domains (Piper et al., 2021; Sang et al., 2022). Recent studies include important applications of this task in supporting professionals in mental health. (Tammewar et al., 2020; Adler et al., 2016; Danieli et al., 2022). Automatic narrative understanding may provide a summarized comprehension of the users' recollections that can be used to engage in personal and grounded dialogues with the narrator. Narrative understanding has been approached in different ways (Kronenfeld, 1978; Chambers and Jurafsky, 2008; Kim and Klinger, 2018). A research direction in this field focuses on extracting the sequence of events that are mentioned in the narrative to obtain a summarized understanding of the whole narrative and its characters (Chen et al., 2021; Mousavi et al., 2021). In these works, the event is mostly represented by a predicate along with its corresponding subject and object dependencies. This definition relies on two assumptions a) the predicate represents an action/occurrence relation between the subject and the object dependencies; b) reoccurring characters across different events are the protagonists of the narrative.

There have been interesting studies on different aspects of events in a narrative such as linking the correlated events as a chain (Chambers and Jurafsky, 2008), learning semantic roles of participants (Chambers and Jurafsky, 2009), commonsense inference (Rashkin et al., 2018), and temporal common-sense reasoning (Zhou et al., 2019).

In order to obtain a concise and salient understanding of the narrative through the events, it is necessary to identify and select the events that relate to a new happening/participant in the narrative and have novel contributions. The process of recognizing a new event implicitly involves the event coreference resolution task, which consists of detecting the mentions of the same event throughout the content (Zeng et al., 2020). Essentially, an event that is referring to a previous event is not considered new. Nevertheless, even if an event appears in the narrative for the first time it might be part of commonsense knowledge, and thus not provide any new information.

In this paper, we address the problem of identifying new events as they unfold in the narrative. This task is inspired and motivated by the need to a) extract salient information in the narrative and position them with respect to the rest of the discourse events and relations, and b) acquire new events from a sequence of sentential units of narratives. This task can facilitate higher levels of computation and interaction such as reasoning, summarization, and human-machine dialogue. Last but not least, we believe this task is a novel and very challenging machine learning task to include in natural language understanding benchmarks.

Proceedings of the The 5th Workshop on Narrative Understanding, pages 1–10
July 14, 2023 ©2023 Association for Computational Linguistics

We assess whether an event is new in a narrative according to their Information Status (IS) (Prince, 1988; Mann and Thompson, 1992). IS refers to whether a piece of information, which can be represented as an entity or other linguistic forms, is new or old. We consider an event new if it has not been previously observed in the context and provides novel information to the reader; that is, its information (the event and/or participants) is not presented priorly in the discourse stretch, and it can not be inferred through commonsense. For instance, *Bob **saw** Alice* is a new event if it is the first time that Alice is introduced in the narrative or the first time Bob saw her. However, once this event is selected as new, *Bob **looked at** Alice* will not be a new event anymore. Furthermore, if *Bob **married** Alice* is considered as a new event, *Alice **is** Bob's wife* can be inferred through commonsense and thus is not a new event. An example of new and old events is presented in Figure 1. While there are eight events in the narrative sentences, two of them do not represent any novel information and thus are not new.

For this purpose, we developed an unsupervised model to extract markable event candidates from the narratives. We parsed a publicly available dataset of narratives, SEND (Ong et al., 2021), and using the developed model, extracted all the markable events for each sentence. In the next step, we designed and conducted an annotation task using five human annotators to select the events in each sentence that are discourse-new with respect to the narrative context. In order to validate the annotation protocol and evaluate the results, we developed several neural and non-neural baselines for the task of new event extraction in both candidate-selection and sequence-tagging settings.

The contributions of this paper can be summarized as follows:

- We present the novel task of new event detection for narrative understanding along with its annotation methodology and evaluation.
- We present the annotated version of a public corpus of emotional narratives for the task of automatic detection of new events in a narrative.
  [1].
- We introduce several baseline benchmarks for the task of new event detection based on discourse heuristics and deep neural networks, in two different settings of candidate selection and sequence tagging.

So uh during my childhood **I had two dogs; one was named Flash**, **one was named Fluff.**

I got them when I was three and around the age of eight, **we were moving to the US** from Guyana.

When we were living in the US, **we rented a house** for a short time and **my father bought a big sofa.**

Figure 1: An example of a narrative and the corresponding events. There are eight events in the sentences (highlighted), while six of them are presenting new information (bold) and the remaining two are referring to the already-mentioned events in the context (not bold).

2 Literature Review

Event Extraction The definition of the event concept has been the topic of study in different disciplines, originating in philosophy (Mourelatos, 1978). Early attempts to understand the semantics and structures of events in the text used hand-coded scripts with predefined slot frames to be filled by the values extracted from the text (Kronenfeld, 1978). This approach was later adopted by other works (Kim and Klinger, 2018; Ebner et al., 2020). Kim and Klinger (2018) consider the activation of emotions as an event and study such events through different properties such as cause, experiencer, target, etc. In this definition, not only verb phrases but also noun phrases and prepositional phrases that manifest an emotion in a narrative participant can represent events. (Ebner et al., 2020) studied the events and their participants by the verb-specific roles the participants can have (the arguments of the event "attack" are of types "attacker" and "target"). In this work, the authors formalized the event understanding as an argument-linking task.

To address the expensive nature of designing domain-specific frames, Chambers and Jurafsky (2008) proposed an unsupervised approach to extract the event chains in a narrative according to the linguistic structures of the narrative sentences. Based on the assumption that reoccurring participants among different events are the protagonists of the narrative, the authors defined an event in a sentence as a predicate (verb) and the verb dependencies including the protagonist. This work was

[1] Link to our Repository

complemented further by considering the role of the protagonists in each event and the neighboring events in order to obtain a schema (Chambers and Jurafsky, 2009).

Event-Centric Understanding There have been several studies on the application of event-centric narrative understanding. Mostafazadeh et al. (2016) studied the understanding of commonsense stories via event chain extraction model (Chambers and Jurafsky, 2008). Rashkin et al. (2018) conducted a task on inferring the next possible intents and reactions of the participants in a narrative based on the observed events through commonsense. Zhou et al. (2019) studied the application of temporal reasoning such as order/frequency of events in the narrative for the question-answering setting. Mousavi et al. (2021) extracted events in a personal narrative to construct the personal space of events and participants in the user's life as a graph.

Event Co-reference Resolution The event coreference resolution task is focused on identifying the events that refer to previously mentioned events in a context. Two events are considered identical if they share the same spatiotemporal location (Quine, 1985). Bejan and Harabagiu (2010) studied the detection of coreferential events by measuring the similarity among two events using lexical and semantic features. Zeng et al. (2020) proposed a model based on BERT pre-trained model (Devlin et al., 2019) to integrate event-specific paraphrases and argument-aware semantic embeddings for this task.

3 Definition of New Event

We introduce the task of identifying the new events in a narrative to obtain a distilled and concise representation of the whole narrative and its characters. We follow the definition of an event that was used by Chambers and Jurafsky (2008) based on the verb and its dependencies. That is, a verb is a core element of an event and supports the relation among its dependencies such as subject, object/oblique nominals which are considered as the participants of the event (Mousavi et al., 2021).

Prince (1988) defined the notion of old or new Information Status (IS) with respect to two aspects of the hearer's beliefs and the discourse model. New information according to the hearer's belief is the one that is assumed not to be already known for the hearer, while discourse-new information is the one that has not been mentioned or has not occurred priorly in the discourse-stretch (Prince, 1988). Nissim et al. (2004) adopts the IS concept and defines three categories of old, new, and mediated for the status of entities in a dialogue. The notion of old follows the definition provided by Prince (1988) closely. However, the authors define mediated as entities that have not been introduced directly in the context but are inferrable or generally known to the hearer; while the new category spans over entities that are not introduced priorly in the dialogue context, nor can they be inferred from the previously mentioned entities.

	Value
#Narratives (Train:Valid:Test)	193 (114:40:39)
#Subject (# female)	49 (30)
Avg. Narrative Len.	28.10 utterances
Avg. Utterance Len.	15.44 tokens
#Vocabulary	4,416 unique tokens

Table 1: The statistics of SEND dataset (Ong et al., 2021). The dataset is provided with official train, valid and test sets. The majority of narrators are female and each narrative consists of approximately 430 tokens on average.

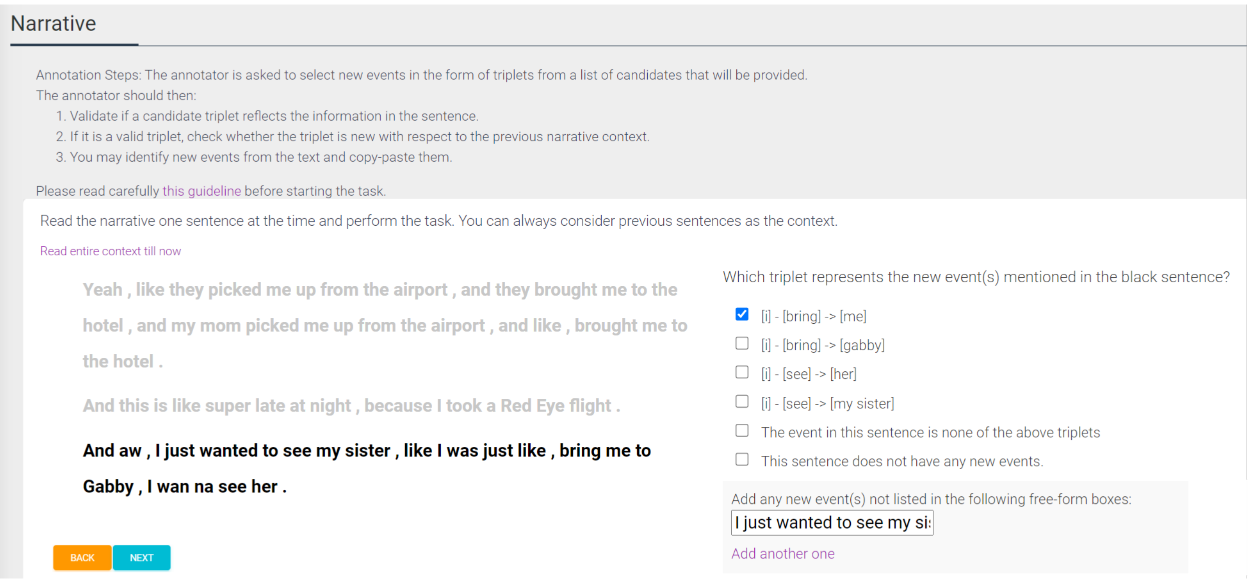
We extend the definition of the new category in entities (Nissim et al., 2004) to events. We define new events as those that are not mentioned in the narrative context and can not be inferred through commonsense by the reader. In this work, we do not consider further distinctions such as old or mediated.

4 Annotation of New Event

4.1 Annotation Task Description

Narrative Dataset We conducted an annotation task for identifying the new events in narratives at the sentence level. The corpus used in this study is the SEND dataset (Ong et al., 2021), which is a collection of emotional narratives. The dataset consists of 193 narratives from 49 subjects, collected by asking each narrator to recount 3 most positive and 3 most negative experiences of her/his life. The statistics of the SEND dataset are presented in Table 1 (the train, valid, and test sets are the official splits).

Task Design To reduce the annotators' workload, we developed a baseline model inspired by Mousavi et al. (2021) to automatically parse and extract all event candidates for each sentence in the narrative as the triplets of (subject, predicate, object). In the cases where more than 5 candidates were extracted for a sentence, we created 5 clus-



Narrative

Annotation Steps: The annotator is asked to select new events in the form of triplets from a list of candidates that will be provided.
The annotator should then:

1. Validate if a candidate triplet reflects the information in the sentence.
2. If it is a valid triplet, check whether the triplet is new with respect to the previous narrative context.
3. You may identify new events from the text and copy-paste them.

Please read carefully this guideline before starting the task.

Read the narrative one sentence at the time and perform the task. You can always consider previous sentences as the context.

Read entire context till now

Yeah , like they picked me up from the airport , and they brought me to the hotel , and my mom picked me up from the airport , and like , brought me to the hotel .

And this is like super late at night , because I took a Red Eye flight .

And aw , I just wanted to see my sister , like I was just like , bring me to Gabby , I wan na see her .

Which triplet represents the new event(s) mentioned in the black sentence?

☑ [i] - [bring] -> [me]
□ [i] - [bring] -> [gabby]
□ [i] - [see] -> [her]
□ [i] - [see] -> [my sister]
□ The event in this sentence is none of the above triplets
□ This sentence does not have any new events.

Add any new event(s) not listed in the following free-form boxes:

I just wanted to see my si:

Add another one

Figure 2: The user interface of the annotation platform. The annotator is presented with the narrative one sentence at a time on the left side of the screen. The event candidates and the option to add new events as free-from text are located on the right side of the interface. Moreover, a short version of the guidelines and the previous context of the narrative are shown to the annotator throughout the annotation.

ters using Levenshtein distance (Yujian and Bo, 2007) (hierarchical clustering) and the candidate with the most number of tokens in each cluster was selected to be presented to the annotator. We randomly sampled 21 narratives from the SEND dataset and reserved them as backup data (13 narratives from the train set, 4 from the valid set, and 4 from the test set). Using the extraction pipeline, we extracted all subject-predicate-object triplets as event candidates in the remaining 172 narratives at the sentence level.

Annotation UI The user interface (UI) of the annotation platform is presented in Figure 2. Throughout the task, the annotator is presented with a brief version of the task guidelines on the top of the display (with access to the complete version). The narrative is presented on the left side of the screen with the current sentence in black and the context in grey. The narrative is updated progressively sentence-by-sentence while the annotator has access to the previous sentences of the context. For each sentence, the annotation question, the list of the triplet candidates and the possibility to select and add continuous span from the text are presented on the right side.

Annotation Task During the task, the annotators were presented with a narrative one sentence at a time and the corresponding list of candidates. They were asked to control if any of the candidate triplets in the list is valid (i.e. it reflects the information in the sentence correctly); and whether it provides new information with respect to the previous narrative context, that can not be inferred through commonsense. In the case of valid and new information, the annotators were asked to select that candidate as a new event. Furthermore, if there were no candidates extracted for a sentence or the new information in a sentence was not presented as a valid candidate, the annotator was asked to add the new information by simply copying the segment that conveys it from the sentence and adding it as continuous span text.

Task Execution We recruited five annotators for the task of new event annotation. The annotators were non-native English speakers with certified English proficiency. After an introductory meeting with the annotators, they were asked to carry out the first qualification task which consisted of annotating one narrative, sampled from the valid set. The result of the first qualification batch was checked manually and a few refinements were made with the annotators. The annotators were then asked to perform a second qualification task using another narrative randomly sampled from the valid set. The Inter-Annotator Agreement (IAA) level during the two qualification tasks, which is presented in Table 2, indicates the improvement in the annotators' performance from one qualification batch to the other. The IAA for the event candidates is calculated using Krippendorff's α (Krippendorff,

Annotation Format	**Qualifications** **First**	 **Second**	**Overall IAA**
Selected Candidates	0.22	0.55	0.54
Added Spans	0.32	0.60	0.66

Table 2: Inter-Annotator Agreement (IAA) during the qualification tasks and over the whole annotation task. The results indicate an improvement in the performance of annotators from one qualification batch to the other. The IAA is computed for candidate selection and continuous span selection annotation using Krippendoff's α and the extension of Cohen's κ for segmentation agreement, respectively.

Sentence 1: So uh during my childhood I had two dogs; one was named Flash, one was named Fluff.

Candidates: a. [i] - [had] -> [my childhood]
b. **[i] - [had] -> [two dogs]** ✔
c. **[one] - [was named] -> [fluff]** ✔
d. [i] - [so had] -> [two dogs]
e. **[one] - [was named] -> [flash]** ✔

Sentence 2: I got them when I was three and around the age of eight we were moving to the US from Guyana.

Candidates: a. [i] - [got] -> [them]
b. **[we] - [were moving tol -> [the us]** ✔
c. [we] - [were moving to] -> [guyana]

Sentence 3: When we were living in the US, we rented a house for a short time and my father bought a big sofa.

Candidates: a. [we] - [were living in] -> [the us]
b. **[we] - [rented] -> [a house]** ✔
Added Spans: ***my father bought a big sofa***

Figure 3: An example of sentences in a narrative and the corresponding events; while the baseline model has extracted various event candidates, only a few of them are valid and new events (bold). Furthermore, the baseline model has missed an event in the third sentence which is added as a span from the sentence.

2011), while the IAA for the continuous span text is calculated by the extension of Cohen's κ for segmentation agreement (Fournier and Inkpen, 2012), averaged among all annotators. The remaining 170 narratives were divided into 11 batches. In each batch, one narrative was annotated by all annotators for the purpose of continuous quality control of the results, while the rest was equally divided among the annotators. To prevent unreliable and biased agreements, all 11 overlapping narratives were from different narrators.

4.2 Annotation Result Evaluation

We annotated the dataset of personal narratives, SEND (Ong et al., 2021), with new events in the sentence level by five human judges. An example of the annotation results is presented in Figure 3. While the baseline model has extracted various possible event candidates from the sentence, only a few of them are **valid** events that are representing ***new*** information. Moreover, the model has failed to extract an event in the third sentence which is added as a span from the text.

Selected New Events as Candidates	
#Candidates selected	1536
Avg. candidates selected:	
per Sentence	0.57
per Narrative	9.0
per Narrator	31.4
%Candidates selected in:	
1st half of the Sentence	43%
2nd half of the Sentence	57%
1st half of the Narrative	55%
2nd half of the Narrative	45%
Added New Events as Continuous Spans	
#Spans added	2254
Avg. spans added:	
per Sentence	0.8
per Narrative	13.3
per Narrator	46.0
%Spans added in:	
1st half of the Sentence	38.1%
2nd half of the Sentence	61.9%
1st half of the Narrative	96.9%
2nd half of the Narrative	3.1%

Table 3: The statistics of the annotated dataset. While only 1536 extracted candidates (out of 6938, thus 22%) were selected as new events, 2254 new events were added by the annotators as continuous span text. Moreover, almost all of the continuous span events appear in the first half of the narrative, while event candidates have a quite normal distribution.

Throughout the task, the IAA level on the overlapping narratives was computed to ensure a consistent annotation quality. We observed negligible fluctuations in the IAA level during the task (<0.9 for Krippendoff's α), except for one batch; for which the low-quality contributions were detected and refinements were made with one annotator. The overall IAA level of the annotated dataset is presented in Table 2. The results are close to the level obtained in the second qualification batch.

The statistics of the annotated dataset, presented in Table 3, indicate that the majority of the annotated events were added as continuous span text and were not extracted by the baseline model. Moreover, while the event candidates appear in the nar-

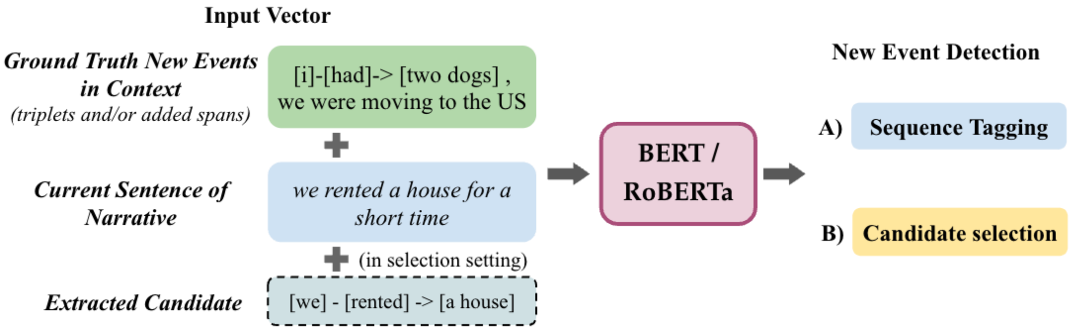

Input Vector
Ground Truth New Events in Context
(triplets and/or added spans)
[i]-[had]-> [two dogs] , we were moving to the US
Current Sentence of Narrative
we rented a house for a short time
(in selection setting)
Extracted Candidate
[we] - [rented] -> [a house]
BERT / RoBERTa
New Event Detection
A) Sequence Tagging
B) Candidate selection

Figure 4: The neural baselines for the task of new event detection. The input vector consists of the new events in the context (ground truth) and the current sentence. In the candidate selection setting, the input vector includes the extracted candidate as an additional segment as well. The model encodes the input vector and outputs either a) a sequence of tags, corresponding to the tokens in the sentence; or b) a binary decision to categorize the candidate as new or not.

	Prec.	Rec.	F1
Random	24.0	29.2	26.3
Binary	22.8	49.4	31.2
First Candidate	27.7	33.7	30.4
Last Candidate	30.1	36.7	33.1
New Subject	24.6	28.6	26.5
New Entity	25.1	88.9	39.1
BERT	35.6	51.1	41.6
RoBERTa	40.4	83.1	**54.3**

Table 4: The results of the new event candidate selection baselines. The performance of the neural models is averaged over 10 runs.

rative with an approximately uniform distribution, almost all of the continuous span events are located in the first half of the narrative. This result is in line with the definition of new events since the events mentioned before in the context are "old" events. Nevertheless, in both cases of candidate events and continuous span events, we observe that the second halves of the sentences contain more information than the other half, indicating that the narrators tend to mention the new events at the end of the sentence.

5 Baselines for New Event Detection

We developed neural and non-neural baselines to validate the outcome of the annotation task, and, as baselines for the novel task of new event detection in a narrative. Considering the two annotation formats of selecting candidates and adding continuous spans, we formalize the task using two settings of candidate selection and sequence tagging.

5.1 Candidate Selection Baselines

The first group of models is tasked to select the new events from the candidates extracted by our baseline model. The rule-based models are:

- **Random Selector**: for each sentence and its event candidates, it randomly picks one candidate as the new event in the sentence.
- **Binary Selector**: for each of the event candidates of a sentence, it randomly decides whether it is a new event or not. Thus, each candidate has a 50% chance of being selected as a new event.
- **First Candidate Selector**: that selects the first event candidate that is extracted for a sentence as the new event.
- **Last Candidate Selector**: which selects the last event candidate that is extracted for a sentence as the new event for the sentence.
- **New Subject Selector**: which selects the first candidate that contains a new (unseen) subject in the list of candidates as the new event. In other words, the number of selected candidates is equal to the number of non-repetitive subjects in the candidate list of the narrative.
- **New Entity Selector**: which selects all the event candidates that include new subjects or new objects at the narrative level. Thus, it selects all candidates unless they differ in the verb only. In that case, it selects one of them as the new event.

Neural Network Models In addition to the rule-based models, we developed neural models based on Pre-trained Language Models (PLMs) as baselines for the task of new event candidate selection presented in Figure 4. For this purpose, we model the input vector with three elements as event candidate, current sentence, and context new events. The context new events denote the new events (ground truth) in the narrative context up to the current sentence. In cases where the size of the input vector exceeds the model limits (for instance 512 tokens per BERT-based models), the model trims the former part of the context new events. The model encodes this vector and outputs the classification decision of whether the event candidate (triplet) is a new event or not. The PLMs we fine-tuned for this purpose are BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019).

The results of the candidate selection baselines are presented in Table 4. We observe that *Last Candidate Selector* has achieved the highest precision level among rule-based models. This is in line with the annotation result analysis, indicating the percentage of selected new event candidates to be slightly higher at the end of sentences. On the other hand, *New Entity Selector* achieves the highest level of recall while having a very low level of precision, as it selects all candidates unless the variation is only in the verb predicate. Moreover, the F1 scores of all the rule-based models are less than 40.0%. This indicates that features such as the novelty in elements or occurrence position are not enough to achieve high performance on the task of new event selection. While both neural models outperform the rule-based ones, RoBERTa outperforms all the baselines in this task by having the highest level of precision while maintaining a high recall.

5.2 Sequence Tagging Baselines

The second group of the models is developed for the task of new event detection in a sequence tagging setting. That is, the models tag the sequence of tokens (chunks) which are representing a new event in the sentence. The analysis performed on the continuous span events selected by the human judges indicated that several events can share the same tag spans such as subject or object. Therefore, we formalize this task as a binary tagging task rather than IOB tagging task and leave the development of the models for IOB tagging of multiple spans with overlap as future work. Similar to the previous task, we developed rule-based and neural baselines for new event sequence tagging. The developed rule-based baselines are:

	Prec. (%)	**Rec.** (%)	**F1** (%)
Random	18.8	49.7	27.3
Early	17.4	29.5	21.9
Late	20.2	34.0	25.4
BERT	33.2	82.2	47.3
RoBERTa	34.3	81.3	**48.3**

Table 5: The results of the new event sequence tagging baselines. The models are trained and tested on continuous span events annotated by the human judges only. The performance of the neural models is averaged over 10 runs.

	Prec. (%)	**Rec.** (%)	**F1** (%)
Random	31.1	49.6	38.2
Early	30.8	31.6	31.2
Late	29.9	30.4	30.2
BERT	54.9	84.3	66.5
RoBERTa	55.5	84.8	**67.1**

Table 6: The results of the new event sequence tagging baselines. Compared to Table 5, in this setting, the models are trained and tested on both selected candidates and continuous span events annotated by the human judges. The performance of the neural models is averaged over 10 runs.

- **Random Tagger**: which randomly tags tokens in a sentence as the new event tokens.
- **Early Tagger**: which tags the tokens in the first 30% of a sentence as the new event tokens.
- **Late Tagger**: which tags the tokens in the last 30% of a sentence as the new event tokens.

Neural Network Models Using BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) PLMs, we developed two neural baselines for this task. The models take as input the current sentence and the context new events which are the sequences of new events in the narrative context up to the current sentence. Similarly to the previous neural baselines, if the input vector exceeds the size limits of the models the former part of the context new events is trimmed. The model encodes this vector and outputs a tag sequence consisting of $\mathbf{E}_{\text{(vent)}}$

or **O**, corresponding to the tokens in the sentence, indicating whether or not they describe a new event.

We initially trained the sequence tagging baselines using the annotated continuous span events. The results of this experiment are presented in Table 5. We observed that precision scores and consequently F1 scores are not significantly different among rule-based models. This indicates that the position of the tokens in the sentence is not the most contributing factor to the prediction accuracy. Similar to the previous task, the neural models have the highest performance among the baselines. However, their precision is considerably lower than the recall.

Similar to the previous task, the neural models have the highest performance among the baselines. However, their performance can be further improved by increasing the precision since it is considerably lower than the recall. The agreement level of the rule-based models is significantly small since the metric takes into consideration the beginning and the end of the tag spans. This is in contrast with the precision and recall metrics which focus on only binary values of each tag.

In the next step, we evaluated the same baseline models using both the selected event candidates and the continuous span annotations as the train and test sets. The results of this experiment, presented in Table 6, show a boost in the performance of all models using the mentioned train and test sets. Nevertheless, the same performance trends among models can be observed in this experiment as well.

6 Conclusions

In this work, we study the events in narratives according to their Information Status. We introduce the new task of identifying new events as they unfold in the narrative. In our definition of the event, the verb is the central element that represents a relation/happening that engages its dependencies such as subject, object, or oblique nominals. Meanwhile, we define an event as new if it provides novel information to the reader with respect to the discourse (discourse-new) and if such information can not be inferred through commonsense. We annotated a complete dataset of personal narratives with new events at the sentence level using human annotators. We then developed several neural and non-neural baselines for the task of new event detection in both settings of candidate selection and sequence tagging. We share the annotated dataset and the baselines with the community. We believe this task can be a novel and challenging task in narrative understanding and can facilitate and support other tasks in natural language understanding, human-machine dialogue, and natural language generation.

7 Limitations

The dataset used in this work is a personal narrative corpus in English collected in-vitro (e.g. subjects in a lab setting). Further work will be needed to extend it to other languages, genres, and naturalistic conditions. The reproducibility of the annotation task may be subject to variability due to the fact that the task is done by five internal annotators and not through crowd-sourcing techniques.

Acknowledgements

The authors would like to thank Desmond C. Ong for providing the dataset and useful discussions about the annotation task.
We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

References

Jonathan M Adler, Jennifer Lodi-Smith, Frederick L Philippe, and Iliane Houle. 2016. The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review*, 20(2):142–175.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021.

Event-centric natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14, Online. Association for Computational Linguistics.

Morena Danieli, Tommaso Ciulli, Seyed Mahed Mousavi, Giorgia Silvestri, Simone Barbato, Lorenzo Di Natale, Giuseppe Riccardi, et al. 2022. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial. *JMIR mental health*, 9(9):e38067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

David B Kronenfeld. 1978. Scripts, plans, goals, and understanding: an inquiry into human knowledge structures by roger c. schank and robert p. abelson. *Language*, 54(3):779–779.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

William C Mann and Sandra A Thompson. 1992. *Discourse description: Diverse linguistic analyses of a fund-raising text*, volume 16. John Benjamins Publishing.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Alexander P. D. Mourelatos. 1978. Events, processes, and states by alexander p. d. mourelatos. *Linguistics and Philosophy*, 2(3):415–434.

Seyed Mahed Mousavi, Roberto Negro, and Giuseppe Riccardi. 2021. An unsupervised approach to extract life-events from personal narratives in the mental health domain. In *Italian Conference on Computational Linguistics 2021 (CLiC-it)*, Milan, Italy.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. 2021. Modeling emotion in complex stories: The stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594.

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.

Ellen F. Prince. 1988. The zpg letter: Subjects, definiteness, and information-status. pages 295–325. John Benjamins.

Willard Van Orman Quine. 1985. Events and reification. *Actions and events: Perspectives on the philosophy of Donald Davidson*, pages 162–171.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022. A survey of machine narrative reading comprehension assessments. *arXiv preprint arXiv:2205.00299*.

Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. 2020. Annotation of emotion carriers in personal narratives. In *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 1517–1525, Marseille, France. European Language Resources Association.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.