

Evaluation and Alignment of Movie Events Extracted via Machine Learning from a Narratological Perspective

Feng Zhou, Federico Pianzola*

Center for Language and Cognition, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

Abstract

We combine distant viewing and close reading to evaluate the usefulness of events extracted via machine learning from audio description of movies. To do this, we manually annotate events from Wikipedia summaries for three movies and align them to ML-extracted events. Our exploration suggests that computational narratology should combine datasets with events extracted from multimodal data sources that take into account both visual and verbal cues when detecting events.

Keywords

movie events, narrative events, computational narratology, audio description, movie summaries

1. Introduction

The events that compose a story are crucial for researchers conducting content analysis of movies [16]. As artificial intelligence aims to achieve the ability to automatically understand narratives as a long-term goal [4], researchers continuously develop and improve machine learning (ML) methods for more accurate event extraction from audiovisual narratives. However, the unstructured nature of video data and advanced semantic content of movies pose challenges for computers in understanding and processing videos [17, 30]. The effectiveness of ML in event extraction from audiovisual material still lags behind human manual extraction [16]. Moreover, narrative understanding is a border and more complex process that involves hermeneutic processes that go beyond the identification of events [21]. In this light, our goal is to evaluate the alignment between human interpretation of the main events in a movie narrative and machine processing of the same narrative. Differently from many works in computer science, our focus is on the understanding of the usefulness of ML methods for narratology.


In this limited exploration, we employed a combined approach of distant viewing and close reading. Firstly, we manually annotated events from Wikipedia movie summaries. Then, we manually aligned them to events automatically extracted from audio descriptions. We specifically focus on coverage of movie content, duration, length, and type of events.


CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

*Corresponding author.

✉ fmarinazhou@gmail.com (F. Zhou); f.pianzola@rug.nl (F. Pianzola)

🆔 0000-0001-6634-121X (F. Pianzola)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Research

Narrative theory defines events as the minimal units of narrative, being states, processes in time, and changes of state [22, 32]. Moreover, narrative events are often analyzed along the dimensions of their duration and frequency in chronological terms, and span of their representation through a medium [10]. In the case of movies, events are represented as scenes, typically taking place at a location, involving a set of characters, and spanning a continuous period of time [8]. In addition, sub-scenes occur within a single location and time frame, are usually shorter, contain several shots that do not constitute a complete semantic unit, and may lack a clear beginning or end [8]. As scenes are sequences of sub-scenes, the shots contained within a series of sub-scenes can be used to create narrative sub-events or whole events [7, 18].

To cope with such granularity, most ML methods extract events through the cross-modal understanding of audio, visual, and video-related text: e.g. alignment between script and time-stamped subtitle [15], retrieval via semantic graphs of video clips and plot summaries [34, 35], segmentation by detecting audio descriptions between character dialogues [31, 23]. The evaluation of ML methods is done via manual annotation of movie events and aligning sentences in movie scripts or summaries with video clips [30, 35, 36]. User-generated content platforms like Wikipedia and IMDb provide a vast amount of movie-related textual data [3, 13] that can provide high-level semantic annotation of movies, such as describing specific scenes and events using text [6], and serve as gold standards to evaluate the ability of machine learning methods for event extraction.

Despite the variety of methods, the relevance of ML-extracted events for summaries that align with human expectations has not been explored in detail. Therefore, we compared a dataset of automatically extracted events to events manually extracted from movie summaries, focusing on alignment, duration, length, and type of events.

3. Data

3.1. Machine Learning Extracted Events from Audio Description

Audio description (AD), also known as Descriptive Video Service (DVS), is conceived for visually impaired audiences and usually contains the most important visual elements of a movie frame, such as scene changes, characters' appearance, actions, and interactions between characters [31]. AD uses short and precise language inserted between characters' dialogue and mixed with the original soundtrack of the movie [31, 23]. AD has been used to extract movie events and research the narrative structure of visual materials because of the high quality of the movie description text and the high degree of alignment with visual content [31]. We used the Montreal Video Annotation Dataset (M-VAD) [31]¹ to evaluate this type of story representation with respect to narrative event granularity and plot coverage [28]. Three movies of

¹Since AD appears between character dialogues, the curators of the M-VAD dataset first separated the AD soundtrack from the movie soundtrack, then they automatically segmented the movie into different events by detecting the pauses between ADs, finally they transcribed the AD of the events into text using a professional transcription service. The data including the movie event description text and the corresponding video clips' timestamps are obtained from the M-VAD Names GitHub repository: <https://github.com/aimagelab/mvad-names-dataset/releases/tag/1.0>. The subset used in this article and the manually-annotated data are available in the repos-

different genres were selected from the M-VAD dataset: *500 Days Of Summer* (2009, romantic comedy, 95 minutes), *The Social Network* (2010, biographical drama, 120 minutes), and *Flight* (2012, thriller drama, 138 minutes).

3.2. Manually Annotated Events from Movie Summaries

Wikipedia crowdsourced summaries can be considered a proxy of the events that people find important in a movie, spanning the whole story arc [28]. They are also used as gold standard in NLP because they have an event granularity that balances efficient plot recognition and information loss [28, 37]. We linked these gold-standard events to the respective video clips by identifying the timestamps corresponding to the events' occurrences in the movie. These events manually extracted from summaries allowed us to evaluate the ML movie event dataset with respect to key content coverage (condensing important parts of information) [29] and the temporal localization of such events in the plot. Additionally, events can be classified by type [9] and used to explore the criteria underlying the summarization process and the plot structure [12]. We annotated events in summary texts based on the main verb in sentences, according to four types: stative, process, change of state in the story world, or none of the previous [32]. Two annotators independently annotated events from Wikipedia summary texts using the software CATMA [11] and then marked the start and end timestamps of the events' corresponding video clips while watching the movies.

4. Methods

To test whether Wikipedia summaries' events cover the semantic content of most video clips describing key events in the movie, we first rescaled between 0 and 1 the timestamp (for video clip index) and annotation's first character's position (for plot sentence index) of all annotated events in the summary text that we were able to align to corresponding video clips. In this way, we obtained a comparable plot sentence index and video clip index [1].² Second, we manually aligned human annotated events with the ML-extracted events based on the respective timestamps to examine whether the ML-extracted events have the same event description granularity as Wikipedia summary events from text and video perspectives [28]. The criterion used for the manual alignment is whether the overlapping time interval between one or several corresponding ML- and human-extracted events is at least 75% of either extracted event(s) duration [24]. Additionally, a human evaluation was conducted to check whether the events extracted via ML can represent the semantics of manually annotated events. Subsequently, to assess the extent of the alignment between ML-extracted and human-annotated events, we counted the number of events, and compared their time duration and distribution throughout the movie [14].

itory: https://github.com/Marina-Zhou/Movie_Event_Alignment

²To avoid altering the depiction of event distribution in the movie summary, the few summary events without a timestamp have not been plotted in 2. Additionally, there are instances where a manually extracted event relates to a number of movie clips. Only the segment with the earliest chronological occurrence is kept, the others are discarded for the sake of clarity of the visualization. The full list of events is anyway available in the released dataset.

5. Results and Discussion

Table 1 shows the raw number of events extracted via ML from audio description and those manually annotated in summaries. These two measurements have to be considered independent because the granularity of the events serves different purposes. In the case of AD, the goal is to be as detailed as possible, whereas in the case of summaries, humans only report important events and can also condense many events with one sentence [20]. The reason why the average number of ML-extracted events aligned to summary events in the movie *The Social Network* is relatively small may be that the dialogue in the movie is dense and there is no pause that allows for AD, as also shown by the lower number of ML-extracted events.

Table 1

Number of events extracted via ML from audio description and by humans from summaries

	500 Days Of Summer (2009)	The Social Network (2010)	Flight (2012)
Number of ML-extracted events from AD	436	260	449
Number of manually-annotated events from summaries	80	43	64
Average number of ML-extracted events aligned to summary events	3.68	1.58	4.00
Inter-annotator agreement (Krippendorff’s alpha)	0.93	0.67	0.83

Figure 1 shows that the duration of ML-extracted events is shorter than that of manually annotated events. The duration of events in summaries varies widely, probably because summaries condense the important events of a movie and a sentence may refer to multiple sub-events. In the movie *Flight*, the summary event [He pilots SouthJet Flight 227 to Atlanta] is aligned to seven consecutive sub-events that occur during the flight. Although the events are closely connected, there are some video segments in the middle that are “progressive scenes composed of sequential, nonrepetitive shots” [16]. Furthermore, in summaries, people can subjectively select events that they find important based on different criteria, sometimes more focused on action, other times more focused on characters [26]. On the other hand, AD is inserted in the limited time between conversations and has therefore a more constant duration [25, 33].

Figure 2 shows for three movies that the manually-annotated WikiPlots events correspond to clips covering almost all time periods in the movie [28] and the normalized plot sentence index and the normalized video clip index in the three movies are roughly linearly correlated. Interestingly, there seems to be a position bias [29] in the summary: all three movies have outliers around one third of the summary (Normalized Plot Sentence Index around 0.4). These events correspond to content towards both the beginning and end of the movie, suggesting that in summaries the timeline of the movie can be altered to provide an overall understanding of the plot. After an event or character is introduced, other related events occurring at different times in the movie may be reported. For instance, the following three sentences from index 0.4 of the summary of the movie *Flight* report events occurring respectively at time 1h15m, 30m, and 1m: [Later, he attends a funeral for Katerina,] [a flight attendant who died in the crash,]

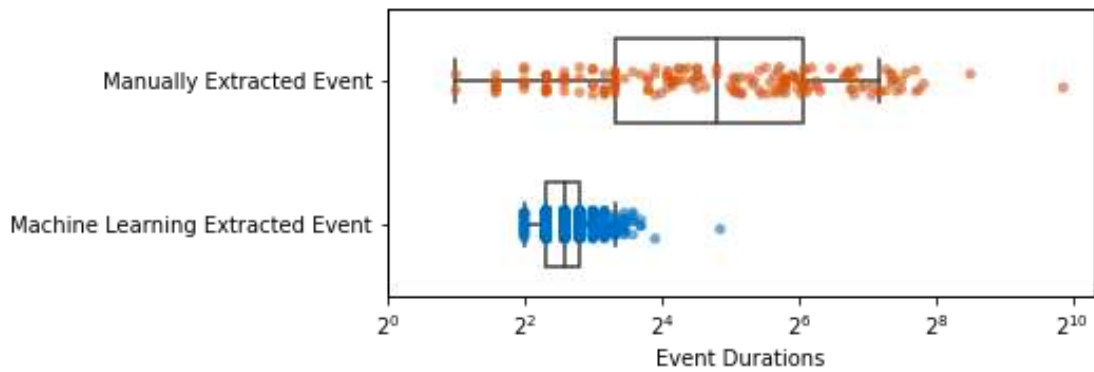


Figure 1: Time duration of manually-annotated (blue) and ML-extracted (yellow) events of three movies

[and with whom Whip had spent the night before the incident].

Another notable phenomenon occurs at the end of the summaries, where many events of the movie ending are reported. The distribution histograms show that the majority of the events in the summaries correspond to events from the last third of the movie. A possible reason is that the end is where usually the most important events of a story occur, sometimes being the confluence of multiple events [5]. But it could also be due to the interpretative function of summaries, which need to provide a sense of closure, therefore the short summary text describes more video clip events at the end of the film to create such effect of closure [27], even though this may not be present in the story.

Figure 3 shows that ML-extracted events cover almost the entire duration of the movie, whereas events in summaries have notable gaps. The width of the bands reflects the length of the reported events and, in Figure 4, it is visible how several ML-extracted events can be encompassed by a longer summary event. There are also some successfully aligned events belonging to sub-scenes about secondary characters that appear in parallel narratives [2, 7].

Figure 5 shows that only part of the ML-extracted events not aligned with summary events occur within the time spans corresponding to manually-annotated events. These events cannot be aligned to a summary event either because the overlapping range with summary events is too small (less than 75%) or because the number of ML-events is insufficient to fully reflect the semantics of the manually-annotated events. For example, in the movie *Flight*, the simple ML-extracted event [Whip turns off the news] overlaps with the much more informative manually-annotated event [he stays with Charlie until the NTSB hearing]. This is due to the mere descriptive function of AD.

The other type of ML-extracted events not aligned with summary events occur outside the time spans covered by summary events. To some extent, these events can be considered progressive scenes that set the background for a summary event and serve as event separators [16]. However, most of them are unimportant events that do not contribute to the plot development. They primarily describe the appearance and secondary actions of the characters, and the visual elements in scenes.

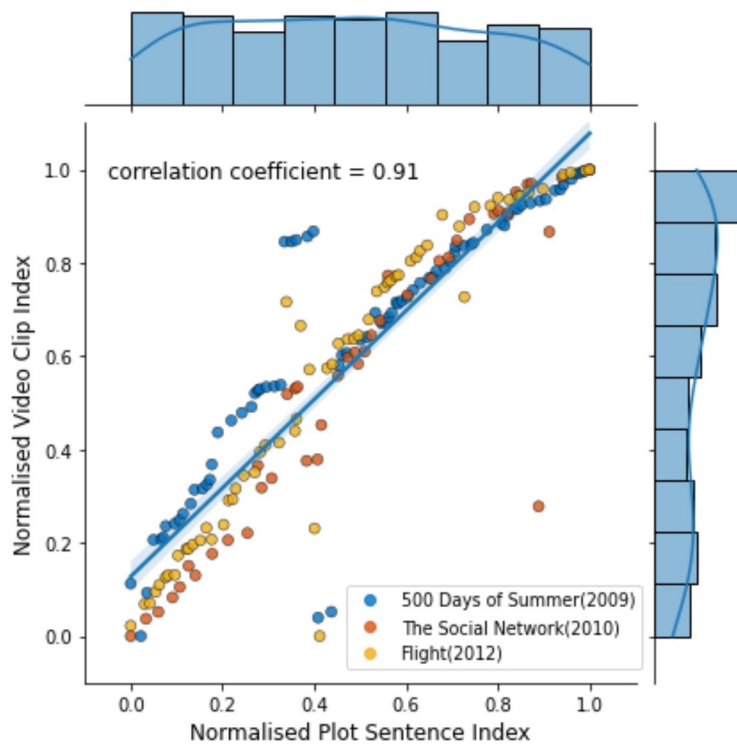


Figure 2: Cumulative distribution of movie summary events and aligned video clips for three sample movies

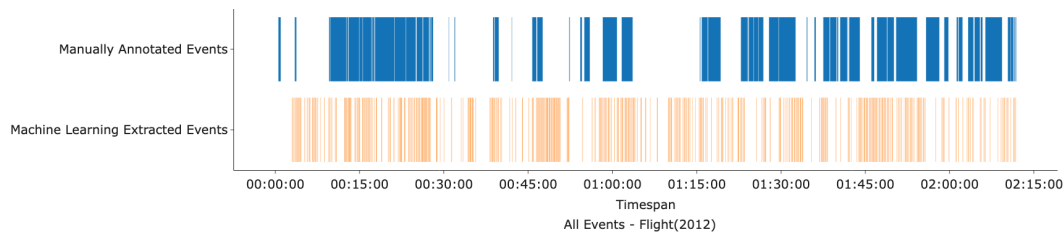


Figure 3: Time length and distribution of all manually-annotated and ML-extracted events for the movie *Flight*

Our annotated data indicates that events extracted from AD more easily align with manually-annotated events when they contain descriptions of visual elements, especially direct descriptions of character actions, rather than information captured through dialogue or background sounds. These ML-extracted events are located within manually-annotated events with longer time duration and may not exhibit direct semantic relevance for such manually-annotated events. Nevertheless, they constitute an integral component of the broader manually-annotated events because they are progressive scenes that separate sub-events. As

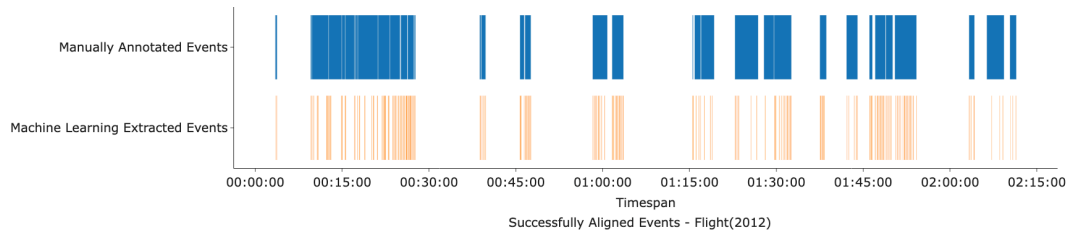


Figure 4: Time length and distribution of manually-annotated events successfully aligned with ML-extracted events for the movie *Flight*

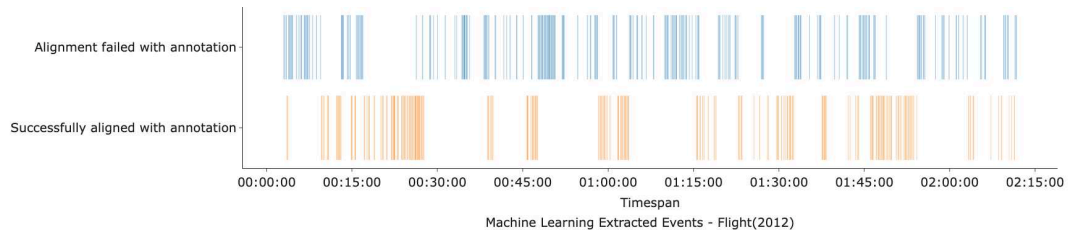


Figure 5: Time length and distribution of ML-extracted events successfully and unsuccessfully aligned with manually-annotated events for the movie *Flight*

such, they have a role in the narrative progression because they introduce new situations where subsequent events happen. For instance, in the movie *Flight*, the ML-extracted event [They reach a clearing, through clouds and bright sunlight, as Whip levels out] serves as a transition, enabling the audience to experience the severe turbulence of the plane and facilitating a smoother acceptance of new information and emotions during a plot transition.

Most of the manually-annotated events are of type “process” (67~78%) and “change of state” (14~16%). Since ML-extracted events contain the description of the characters’ actions, it is easier to align them with manually-annotated “process” and “change of state” events, than with “stative” events (see Figure 6). Notably, most of the “change of state” events in the summary are aligned with ML-extracted events, suggesting that these are indeed important plot turns [19]. However, we were not able to align some “change of state” events from the movies *500 Days of Summer* and *The Social Network* (see Appendix A). This suggests that AD events could still miss information that is important for the plot.

6. Limitations

- The dataset used in this research only contains data from three different movies. Therefore, clear generalization about the usefulness of events extracted from AD via ML for the understanding of narrative cannot be drawn.
- The source of ML-extracted events are AD scripts, which are written by humans. Therefore, this type of data cannot be considered completely machine-generated. A comparison with captions generated by multimodal Large Language Models could provide a

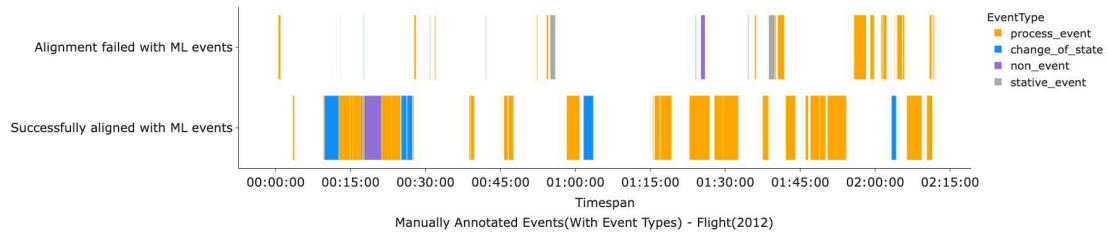


Figure 6: Event type comparison between manually-annotated events successfully and unsuccessfully aligned with ML-extracted events for the movie *Flight*

better case study.

- AD only describes the visual elements of the movie, not the events performed through the character’s dialogue, e.g someone angrily yelling at someone else could be a highly relevant event, maybe a break-up between two partners. Therefore, we suggest to combine AD with data sourced from subtitles or movie scripts.
- The amount of manually-annotated data is quite small (only three movies), so for all the analyses we retained all data annotated by one annotator. The inter-annotator agreement is reported only for reference. Therefore, some subjectivity is unavoidable.
- There is a 1-2 seconds delay between ADs and the corresponding scenes in the movie, so the curators of the M-VAD dataset added 2 seconds at the end of the video to compensate for this deviation [31]. However, this method may also cause inaccurate timestamps. For example, [He takes Whip’s hand] in the movie *Flight* has the end timestamp 01:26:56 but the event actually ends 01:26:54. Moreover, in the M-VAD dataset, the first 3 minutes of the movie are discarded. All these choices slightly affect the alignment with our manually-annotated events.
- We removed some manually-annotated events not corresponding to any video clip in the movie when plotting the visualization, e.g. an event reported by a character but not shown on screen.

7. Conclusions

On one hand, most of the events extracted by machine learning from AD can be considered sub-events rather than narrative events. Because their time duration is very short, they cannot be considered as complete semantic units and need to be understood in conjunction with the video context and dialogue. These events often include descriptions of visual elements in the film and are placed between character dialogues. Some of these events serve as progressive scenes, both inside and outside the automatically extracted event, and are used as separators for other sub-events and events. On the other hand, the events extracted manually from the summary are a high-level synthesis of the dialogue and actions of the characters in the movie, and cover longer time spans that may contain multiple sequential events and sub-events. Moreover, verbs in summary events often refer to one of the four categories of narratological events and are highly relevant for the understanding of the plot, but some verbs in AD events describe

the settings and cannot be considered narratological events.

Computational narratology can certainly make use of ML-extracted events based on audio description but there are strong limitations that suggest using them in combination with other data sources that take into account both visual and verbal cues when detecting events. The multimodal communicative nature of movies seems to pose a big challenge for computational narratology because existing datasets used for ML can only provide an extremely simplified notion of narrative event.

In order to train ML models to extract narrative events, there is a need for movie events datasets that include audiovisual information, subtitles, and AD sentences. The annotation guidelines should also take into account that the narratological concept of event, as unit of a story, should also be related to other conceptualization of events, namely sub-events as those observed in AD and macro-events as those observed in movie summaries.

Since events and causal relationship between events play a key role in narrative understanding [28], our multimodal human evaluation method on the M-VAD movie event dataset should be replicated to evaluate the ability of other automatic methods for the extraction of narrative events. This could help to understand how summary events are constructed by smaller units such as events displayed in the story and sub-events.

In addition, we have not yet conducted research on different types of events extracted by ML and the story arcs that they form. In the future, we could also explore whether movie events extracted via different ML methods contain important narrative events (key plot turns) in the story arc [12].

References

- [1] M. Bain, A. Nagrani, A. Brown, and A. Zisserman. “Condensed Movies: Story Based Retrieval with Contextual Embeddings”. In: *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 – December 4, 2020, Revised Selected Papers, Part V*. Kyoto, Japan: Springer-Verlag, 2020, pp. 460–479. doi: 10.1007/978-3-030-69541-5_28.
- [2] D. Bordwell, K. Thompson, and J. Smith. *Film art: An introduction*. Vol. 8. McGraw-Hill New York, 2008.
- [3] M. Burghardt, A. Heftberger, J. Pause, N.-O. Walkowski, and M. Zeppelzauer. “Film and video analysis in the digital humanities—an interdisciplinary dialog”. In: *Digital Humanities Quarterly* 14.4 (2020). URL: <http://digitalhumanities.org:8081/dhq/vol/14/4/000532/000532.html>.
- [4] S. Chaturvedi, S. Srivastava, and D. Roth. “Where Have I Heard This Story Before? Identifying Narrative Similarity in Movie Remakes”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 673–678. doi: <https://doi.org/10.18653/v1/N18-2106>.

- [5] N. Cohn. “Visual Narrative Structure”. In: *Cognitive Science* 37.3 (2013), pp. 413–452. DOI: <https://doi.org/10.1111/cogs.12016>.
- [6] A. Cooper, F. Nascimento, and D. Francis. “Exploring Film Language with a Digital Analysis Tool: the Case of Kinolab.” In: *DHQ: Digital Humanities Quarterly* 15.1 (2021). URL: <https://www.digitalhumanities.org/dhq/vol/15/1/000515/000515.html>.
- [7] J. E. Cutting. “Event segmentation and seven types of narrative discontinuity in popular movies”. In: *Acta Psychologica* 149 (2014), pp. 69–77. DOI: <https://doi.org/10.1016/j.actpsy.2014.03.003>.
- [8] J. E. Cutting. “Narrative theory and the dynamics of popular movies”. In: *Psychonomic Bulletin & Review* 23.6 (2016), pp. 1713–1743. DOI: <https://doi.org/10.3758/s13423-016-1051-4>.
- [9] S. Dunn and M. Schumacher. “Explaining Events to Computers: Critical Quantification, Multiplicity and Narratives in Cultural Heritage.” In: *DHQ: Digital Humanities Quarterly* 10.3 (2016). URL: <http://www.digitalhumanities.org/dhq/vol/10/3/000262/000262.html>.
- [10] G. Genette. *Narrative discourse: an essay in method*. Ithaca, N.Y: Cornell University Press, 1980.
- [11] E. Gius, J. C. Meister, M. Meister, M. Petris, C. Bruck, J. Jacke, M. Schumacher, D. Gerstorfer, M. Flüh, and J. Horstmann. *Catma*. 2022. DOI: 10.5281/zenodo.6419805.
- [12] E. Gius and M. Vauth. “Towards an Event Based Plot Model. A Computational Narratology Approach”. In: *Journal of Computational Literary Studies* 1.1 (2022). DOI: <https://doi.org/10.48694/jcls.110>.
- [13] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. “On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges”. In: *Eurographics Conference on Visualization (EuroVis) - STARS* (2015). Ed. by R. Borgo, F. Ganovelli, and I. Viola, pp. 83–103. DOI: 10.2312/eurovisstar.20151113.
- [14] N. W. Kim, B. Bach, H. Im, S. Schriber, M. Gross, and H. Pfister. “Visualizing Nonlinear Narratives with Story Curves”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 595–604. DOI: <https://doi.org/10.1109/TVCG.2017.2744118>.
- [15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. “Learning realistic human actions from movies”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, AK, USA: Ieee, 2008, pp. 1–8. DOI: <https://doi.org/10.1109/CVPR.2008.4587756>.
- [16] Y. Li, S. Narayanan, and C. Kuo. “Content-Based Movie Analysis and Indexing Based on AudioVisual Cues”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14.8 (2004), pp. 1073–1085. DOI: <https://doi.org/10.1109/TCSVT.2004.831968>.
- [17] C. Liu, A. Shmilovici, and M. Last. “MND: A New Dataset and Benchmark of Movie Scenes Classified by Their Narrative Function”. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by L. Karlinsky, T. Michaeli, and K. Nishino. Vol. 13804. Cham: Springer Nature Switzerland, 2023, pp. 610–626. DOI: https://doi.org/10.1007/978-3-031-25069-9_39.

- [18] J. P. Magliano and J. M. Zacks. “The Impact of Continuity Editing in Narrative Film on Event Segmentation: Cognitive Science”. In: *Cognitive Science* 35.8 (2011), pp. 1489–1517. doi: <https://doi.org/10.1111/j.1551-6709.2011.01202.x>.
- [19] P. Papalampidi, F. Keller, and M. Lapata. “Movie Plot Analysis via Turning Point Identification”. 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1707–1717. doi: 10.18653/v1/D19-1180.
- [20] V. F. Perkins. “Where is the world? The horizon of events in movie fiction”. In: *Style and meaning: Studies in the detailed analysis of film* (2005), pp. 16–41.
- [21] J. Phelan and P. J. Rabinowitz, eds. *Understanding narrative*. The theory and interpretation of narrative series. Columbus: Ohio State University Press, 1994.
- [22] G. Prince. *A Grammar of Stories: An Introduction*. Berlin, Boston: De Gruyter, 1974. doi: <https://doi.org/10.1515/9783110815900>.
- [23] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. “A Dataset for Movie Description”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3202–3212. doi: <https://doi.org/10.1109/CVPR.2015.7298940>.
- [24] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. “Movie Description”. In: *International Journal of Computer Vision* 123.1 (2017), pp. 94–120. doi: <https://doi.org/10.1007/s11263-016-0987-1>.
- [25] A. Salway. “A corpus-based analysis of audio description”. In: Leiden, The Netherlands: Brill, 2007, pp. 151–174. doi: https://doi.org/10.1163/9789401209564_012.
- [26] J. Sang and C. Xu. “Character-based movie summarization”. In: *Proceedings of the 18th ACM international conference on Multimedia*. Firenze Italy: Acm, 2010, pp. 855–858. doi: <https://doi.org/10.1145/1873951.1874096>.
- [27] E. Segal. *Beginnings and Endings*. 2019. doi: <https://doi.org/10.1093/acrefore/9780190201098.013.1051>.
- [28] Y. Sun, Q. Chao, Y. Ji, and B. Li. *Synopses of Movie Narratives: a Video-Language Dataset for Story Understanding*. 2022. url: <http://arxiv.org/abs/2203.05711>.
- [29] S. Syed, T. Yousef, K. Al Khatib, S. Jänicke, and M. Potthast. “Summary Explorer: Visualizing the State of the Art in Text Summarization”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 185–194. doi: <https://doi.org/10.18653/v1/2021.emnlp-demo.22>.
- [30] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. “MovieQA: Understanding Stories in Movies through Question-Answering”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4631–4640. doi: <https://doi.org/10.1109/CVPR.2016.501>.
- [31] A. Torabi, C. Pal, H. Larochelle, and A. Courville. *Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research*. 2015. arXiv: 1503.01070 [cs.CV].

- [32] M. Vauth, H. O. Hatzel, E. Gius, and C. Biemann. “Automated Event Annotation in Literary Texts.” In: *Chr.* 2021, pp. 333–345. URL: <https://ceur-ws.org/Vol-2989/short%5C%5Fpaper18.pdf>.
- [33] G. Vercauteren. “A narratological approach to content selection in audio description: towards a strategy for the description of narratological time”. In: *MonTI. Monografías de Traducción e Interpretación* 4 (2012), pp. 207–231. DOI: <https://doi.org/10.6035/MonTI.2012.4.9>.
- [34] P. Vicol, M. Tapaswi, L. Castrejon, and S. Fidler. “MovieGraphs: Towards Understanding Human-Centric Situations from Videos”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8581–8590. DOI: <https://doi.org/10.1109/CVPR.2018.00895>.
- [35] Y. Xiong, Q. Huang, L. Guo, H. Zhou, B. Zhou, and D. Lin. “A Graph-Based Framework to Bridge Movies and Synopses”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 4591–4600. DOI: <https://doi.org/10.1109/ICCV.2019.00469>.
- [36] Q. Yi, G. Zhang, J. Liu, and S. Zhang. “Movie Scene Event Extraction with Graph Attention Network Based on Argument Correlation Information”. In: *Sensors* 23.4 (2023), p. 2285. DOI: <https://doi.org/10.3390/s23042285>.
- [37] J. M. Zacks and B. Tversky. “Event structure in perception and conception.” In: *Psychological Bulletin* 127.1 (2001), pp. 3–21. DOI: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.127.1.3>.

A. Appendix

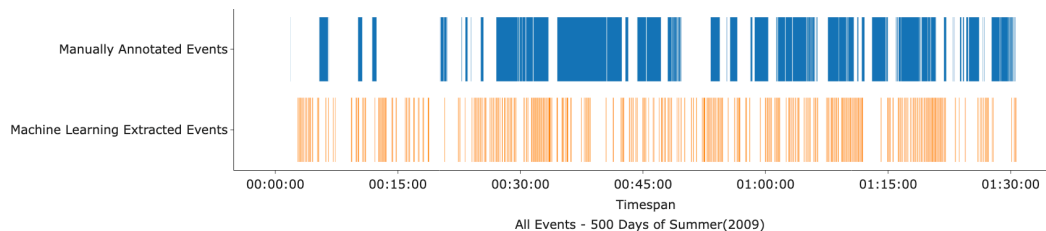


Figure 7: Time length and distribution of all manually-annotated and ML-extracted events for the movie *500 Days of Summer*

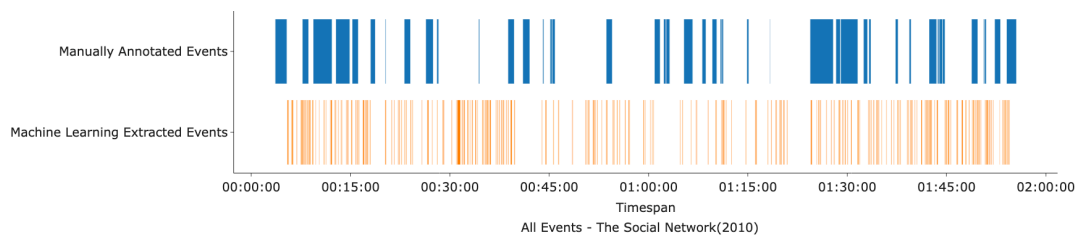


Figure 8: Time length and distribution of all manually-annotated and ML-extracted events for the movie *The Social Network*

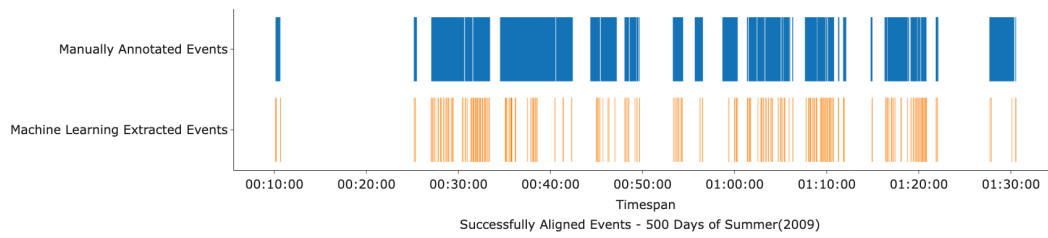


Figure 9: Time length and distribution of manually-annotated events successfully aligned with ML-extracted events for the movie *500 Days of Summer*

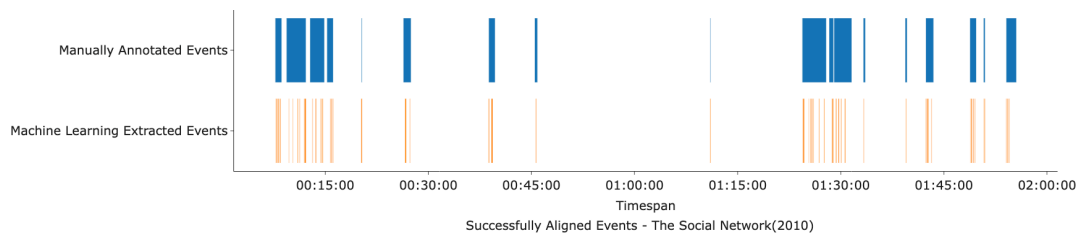


Figure 10: Time length and distribution of manually-annotated events successfully aligned with ML-extracted events for the movie *The Social Network*

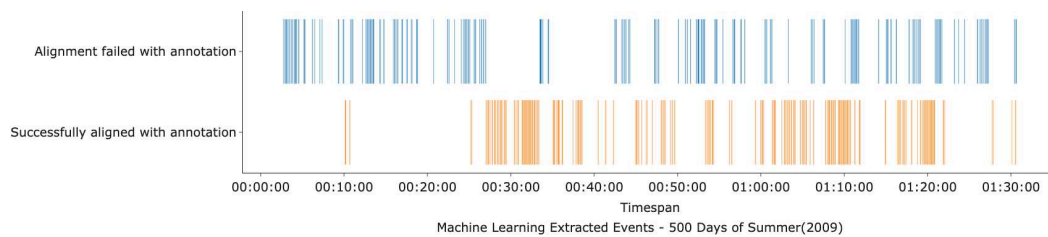


Figure 11: Time length and distribution of ML-extracted events successfully and unsuccessfully aligned with manually-annotated events for the movie *500 Days of Summer*

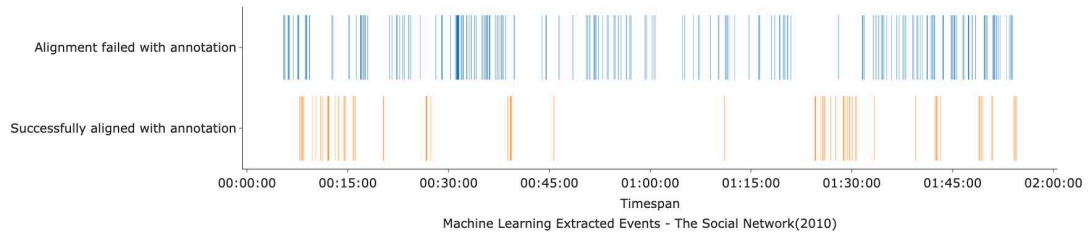


Figure 12: Time length and distribution of ML-extracted events successfully and unsuccessfully aligned with manually-annotated events for the movie *The Social Network*

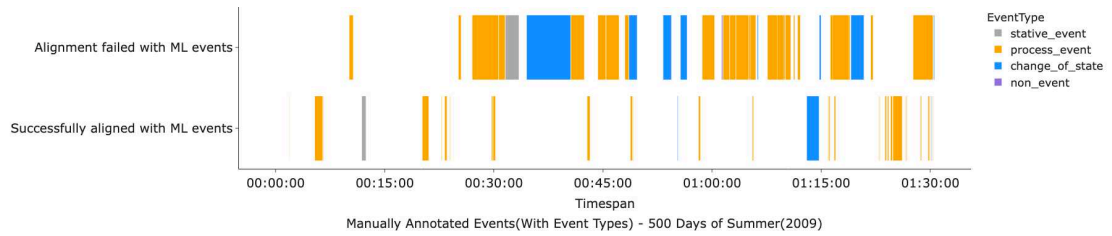


Figure 13: Event type comparison between manually-annotated events successfully and unsuccessfully aligned with ML-extracted events for the movie *500 Days of Summer*

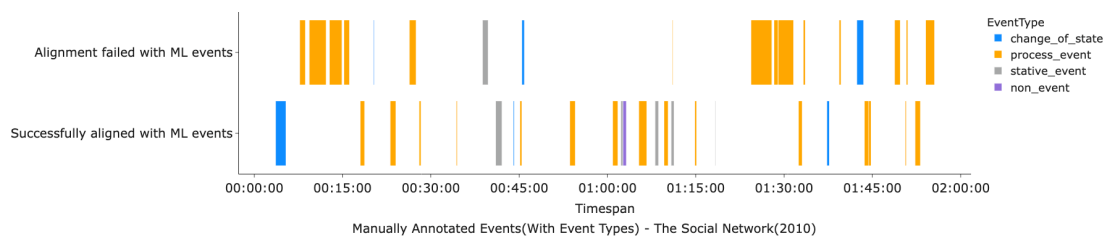


Figure 14: Event type comparison between manually-annotated events successfully and unsuccessfully aligned with ML-extracted events for the movie *The Social Network*