The Inherence of Telicity: Unveiling Temporal Reasoning in Video Question Answering

Olga Loginova[1], Raffaella Bernardi[2]

[1]DISI, University of Trento, Via Sommarive, 9, 38123 Povo TN, Italy
[2]CIMeC, University of Trento, Corso Bettini 31, 38068 Rovereto TN, Italy

Abstract

Video question answering (VQA) requires models to understand video-related questions and generate natural language answers. In multiple-choice VQA, models must associate visual content with one of several predetermined answers. As videos often encompass intricate events and actions unfolding over time, these models must possess the ability to reason across multiple frames and discern the relationships between them with respect to the answers. This paper focuses on the Answerer component of a multiple-choice VQA model, which predicts answers using language-infused key frames. We hypothesise that the Answerer's capacity for temporal reasoning is closely intertwined with its understanding of aspectuality. To investigate this, we augment NeXT-QA, a VQA dataset for causal and temporal reasoning, with annotations for telicity. We then delve into the performance evaluation of SeViLA, a state-of-the-art multiple-choice VQA model, on it. Our findings demonstrate that the model generally exhibits correct handling of aspects, albeit with a bias that is inherent in human nature.

Keywords

video question answering, temporal reasoning, aspect, telicity

1. Introduction

Temporal ordering of actions and events is not solely determined by time; it is also influenced by causality. The organisation of activities in episodic memory is established based on contingency, where one activity triggers another [1]. Recognising cause-effect relationships is essential for temporal understanding, as causes typically precede effects. A cause that has reached its culmination induces the effect.

In language, linguistic aspects play a role in how activities unfold and whether they have culminated. The concept of telicity marks the endpoint of an activity: a verb phrase with a clear endpoint is considered telic (e. g., "to pick up something"), while an atelic one is ongoing, without a specific endpoint (e. g., "to clap"). In descriptions of a sequence of activities with the resultative structure there is an evident human bias towards telic interpretation [2].

Previous research explored telicity for textual transformer-based [3] models, showing that they can classify activities based on duration and telicity with an accuracy surpassing 80% [4]. Such performance at a level comparable to humans, even with limited training data, indicates their ability to capture temporal reasoning through aspect classification.

Our work extends this line of research to video-language models, where video content comes with text labels assigned to key frames or the whole video. Ordering of events corresponds to changing frames, making the correct key frame extraction critical for temporal reasoning. Action timestamps to the frames provide additional cues for temporal reasoning. We propose a study that focuses on contemporary video question-answering (VQA) models in order to explore the relevance of telicity for answering temporal questions related to simultaneous and consecutive activities. We consider the aspects of question's both main and dependent clauses.

To achieve this, we annotate[1] the test set of NExT-QA [5], widely used for causal and temporal reasoning benchmarks, with telicity and evaluate the SeViLA model [6] on this annotated dataset. To the best of our knowledge, this is the first such endeavor in the VQA field.

Q: what is the girl doing with her hands (A)
as the man moved his head sideways at the end of the video (T)

a0: smiles (A)	a1: jamming (A)	a2: pick up something (T)	a3: clap (A)	**a4: pick up toy (T)**

Figure 1: Example of a temporal question and answer options in NexT-QA augmented with our annotation for telic (T) and atelic (A) actions.

[1]The annotations for the dataset are publicly available on GitHub: https://github.com/ologin/Telicity-on-NExT-QA.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
olga.loginova@unitn.it (O. Loginova);
raffaella.bernardi@unitn.it (R. Bernardi)
https://github.com/ologin (O. Loginova);
http://disi.unitn.it/~bernardi/ (R. Bernardi)
0009-0006-1885-3759 (O. Loginova); 0000-0002-3423-1208 (R. Bernardi)

Figure 1 provides an example of NExT-QA temporal questions with our telicity annotation: SeViLA selects the telic answer “pick up toy" (in bold) that does not match the ongoing nature of the question’s main clause “is ... doing", while in the correct answer “clap" (boxed) there is a match of atelic activities.

Our findings demonstrate that the VQA model SeViLA can effectively handle telicity. Furthermore, when making a mistake in prediction the model, like humans, tends to adopt a telic-prone approach.

2. Related Literature

Numerous transformer-based models tackle the challenge of video question answering [7, 8, 9, 10, 11, 12, 6, 13, 14]. These models process both the visual and textual modalities by incorporating video, captions or subtitles, and fuse these streams to generate the final answer. They have showed impressive performance in modelling multimodal VQA. However, they were never assessed for telicity. SeViLA [6], selected for our experiment, consists of two modules: Localizer, for action recognition within videos, and Answerer. The modules are fine-tuned based on BLIP-2 [15]. The model has proved the best results in comparison to other similar models on several datasets, such as STAR [16], NExT-QA [5], How2QA [17], and TVQA [18].

We examined datasets that offer multiple-choice answer options where models must choose the correct answer from a set of candidates. CausalQA [19], Social-IQ [20], CLEVRER [21], STAR [16], and NExT-QA [5] are specifically designed to explore temporal dynamics and the role of causal relationships. NExT-QA proved to be particularly suitable for our experiment, as it is the most comprehensive and emphasises the real-world scenarios.

3. Annotation

The NExT-QA test set comprises 1000 videos with 8564 question-answer pairs supported by five answer options each. From a range of 1 to 15 questions with an average of 9-10 questions per video, we selected solely temporal (T-type) questions. We further excluded closed questions and questions that do not involve two distinct temporally-linked activities, such as “did the baby get hurt after putting out the candle" or “what are the people in this video doing". Thus, the refined total set (RTS) consists of 2060 question-answer pairs.

Notably, RTS questions pertaining to the following activities are in the absolute majority, while the ones concerning preceding actions are very few.[2]

3.1. Aspect Annotation

We divided all question activities into two groups: activities of the main clause (MCA) and activities of the dependent clause (DCA). We annotated independently both question groups, as well as the target and predicted answers, with the following labels of the internal temporal structure:

- **T (telic)** for activities implying an endpoint (e. g., “what happened", “pick up camera", “after the door opens"),
- **A (atelic)** for enduring processes (e. g., “how is the person in black positioned", “smiles", “while watching"), and
- **U (undefined)** for activities lacking clear telicity and duration (e. g., “what does the dog do", “do the same", “to man’s action to him").

Additionally, an **I (irrelevant)** marker was assigned to answers unrelated to aspectuality, such as “astonished" or “nothing". This marker appears among the target answers too in response to questions like “how did the boy react to..." or “what does the person do while...".

Table 1
Telicity of all activities in RTS: questions’ main and dependent clauses, as well as the correct (target) answers

Activity Group	T	A	U	I
MCA	40	159	1861	0
DCA	1254	801	5	0
Answer (target)	758	1283	0	19

From Table 1 it is evident that the question’s main clause rarely impose a definitive telic label, setting the model free to explore temporal relations without predefined constraints. The majority of DCAs are telic and, considering that the most of RTS questions center around the following activity, this affirms the cause-effect nature of the dataset, where the cause predominantly culminates in an endpoint.

4. Experiment and Results

We ran zero-shot SeViLA setting on the test dataset decreasing the batch size down to 2. The obtained results revealed the overall accuracy of 63.18% and the T-type question accuracy of 60,18%. On RTS, we obtained 58.1% of matching predicted and target answers.

We further calculated the telicity precision, recall, F1 score and accuracy on the annotated RTS.

[2]More details on the dataset are in Section A of the Appendix.

4.1. Results

SeViLA selected 781 telic (T) and 1261 atelic (A) responses, alongside 2 instances marked as undefined (U), and 16 responses classified as irrelevant (I). [3]

As demonstrated in Table 2, the results verify that the model attains an accuracy rate exceeding 80%.

Table 2
Telicity precision, recall, F1 score and accuracy results on RTS

Metric	Value
Precision	0.76
Recall	0.74
F1 score	0.75
Accuracy	0.81

The confusion matrix shown in Figure 2 indicates a higher frequency of atelic answers. The majority of atelic responses might initially prompt an inference of an atelic predisposition of the model. Upon closer examination, however, we observed that the incidence of erroneous allocations from atelic to telic responses is more pronounced than in the inverse direction. Thus, the model exhibits a clear inclination towards selecting telic values instead of the target atelic ones: in 26,12% of the target atelic answers it chooses the telic ones, while there are only 14.45% of the opposite cases.

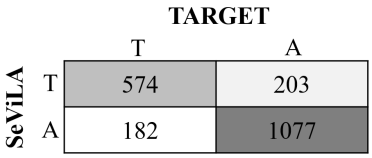

Figure 2: Confusion matrix for telicity classification in RTS. U and I labels are excluded as uninformative.

4.2. Qualitative analysis

The SeViLA Answerer employs a top-k frame extraction strategy to evaluate each frame's probability and determine the optimal choice for answering a question. The erroneous answers often come from the model's misjudgment in instructive key frames.

As shown in Figure 3, the telicity cues may have their origins in the question's both MCA and DCA. As much as in the TN-question (top) SeViLA disregards the DCA's telic action, it also struggles to correspond with the atelic activities of the MCA in the answer for the TC-question (bottom).

[3]Additional data regarding SeViLA's predictions in the context of RTS can be found in Section B of the Appendix.

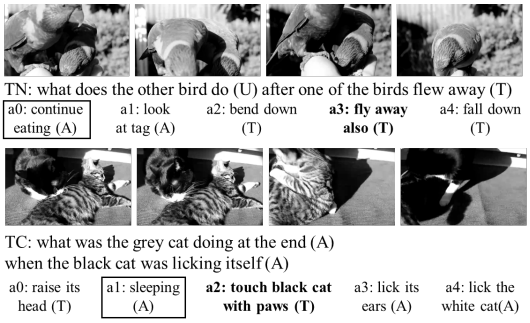

Figure 3: Instances of the key frame misjudgement for next (TN) and current (TC) activities: the SeViLA answers (in bold) and the target ones (boxed).

5. Limitations

While NExT-QA is distinguished as a versatile dataset, it has limitations in representing temporal expressions from a linguistic perspective. Primarily, its questions use a limited set of temporal conjunctions, including *after*, *before*, *during*, *as*, *while*, and *whenever*. A dataset with a broader array of temporal constructions related to both time and telicity could introduce variations, potentially altering model's outcomes.

Another source of result variations can stem from the number of annotators. The annotations were created by a professional linguist in a pilot version, but it is important to acknowledge a potential subjective bias. To mitigate the bias, at least three annotators are suggested for each question-answer pair.

6. Conclusion

The linguistic models grounded in cognitive research highlight a tendency for individuals to remember causally linked activities. Sequential actions and events are associated with the idea that the culmination of one activity sets off another. This culmination is closely tied to the internal structure of the activity which is expressed in language through aspects and, in particular, telicity.

Using NExT-QA dataset, we revealed that VQA models, such as SeViLA, generally capture the contrast in durative and endpoint activities at a human level. Whereas they mostly tend to predict correct telicity for causal and temporal reasoning, their inherent erroneous implication of culminated activity, in essence, aligns with human intuition.

This revelation prompts us to answer the follow-up question: to what extent the improvement in matching telicity in questions and answers will amplify the key frame extraction for correct answering in multiple-choice VQA models.

Acknowledgments

The second author extends her gratitude to Amazon Alexa for their research donation, which significantly supported her work.

References

[1] M. Moens, M. Steedman, Temporal ontology and temporal reference, Comput. Linguist. 14 (1988) 15–28.

[2] Y. Zhao, J. G. Ngui, L. Hall Hartley, S. Bethard, Do pretrained transformers infer telicity like humans?, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 72–81. doi:10.18653/v1/2021.conll-1.6.

[3] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017.

[4] E. Metheniti, T. Van De Cruys, N. Hathout, About time: Do transformers learn temporal verbal aspect?, in: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 88–101. doi:10.18653/v1/2022.cmcl-1.10.

[5] J. Xiao, X. Shang, A. Yao, T.-S. Chua, Next-qa: Next phase of question-answering to explaining temporal actions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9777–9786.

[6] S. Yu, J. Cho, P. Yadav, M. Bansal, Self-chained image-language model for video localization and question answering, ArXiv abs/2305.06988 (2023).

[7] L. Zhu, Y. Yang, Actbert: Learning global-local video-text representations, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 8743–8752.

[8] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, H. Takemura, Bert representations for video question answering, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1545–1554. doi:10.1109/WACV45572.2020.9093596.

[9] A. U. Khan, A. Mazaheri, N. da Vitoria Lobo, M. Shah, Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering, in: Findings, 2020.

[10] J. Lei, L. Yu, T. L. Berg, M. Bansal, Tvqa+: Spatio-temporal grounding for video question answering, ArXiv abs/1904.11574 (2019).

[11] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, Y. Choi, Merlot: Multimodal neural script knowledge models, in: Neural Information Processing Systems, 2021.

[12] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, T.-S. Chua, Video question answering: Datasets, algorithms and challenges, ArXiv abs/2203.01225 (2022).

[13] J. Xiao, P. Zhou, T.-S. Chua, S. Yan, Video graph transformer for video question answering, in: European Conference on Computer Vision, 2022.

[14] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, T.-S. Chua, Contrastive video question answering via video graph transformer, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 13265–13280.

[15] J. Li, D. Li, S. Savarese, S. C. H. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, ArXiv abs/2301.12597 (2023).

[16] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, C. Gan, Star: A benchmark for situated reasoning in real-world videos, in: Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS), 2021.

[17] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, J. Liu, Hero: Hierarchical encoder for video+language omni-representation pre-training, in: Conference on Empirical Methods in Natural Language Processing, 2020.

[18] J. Lei, L. Yu, M. Bansal, T. L. Berg, Tvqa: Localized, compositional video question answering, in: Conference on Empirical Methods in Natural Language Processing, 2018.

[19] A. Bondarenko, M. Wolska, S. Heindorf, L. Blübaum, A.-C. Ngonga Ngomo, B. Stein, P. Braslavski, M. Hagen, M. Potthast, CausalQA: A benchmark for causal question answering, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3296–3308.

[20] A. Zadeh, M. Chan, P. P. Liang, E. Tong, L.-P. Morency, Social-iq: A question answering benchmark for artificial social intelligence, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8799–8809. doi:10.1109/CVPR.2019.00901.

[21] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, J. B. Tenenbaum, Clevrer: Collision events for video representation and reasoning, in: International Conference on Learning Representations, 2020.

A. NExT-QA RTS Dataset Statistics

This section provides more details on the dataset used for the experiment.

A.1. Types of Questions

There is prominent imbalance among all T-type questions in RTS with the majority of TN questions.

Table 3
All types of temporal questions in RTS: questions that ask previous (TP), next (TN) and current (TC) activities

Type	Number	Percentage
TP	91	4,42
TN	1333	64,71
TC	636	30,87

A.2. Question Structure

The detailed overview of RTS questions shows that the dataset has predominately questions with the “what did S do after..." structure.

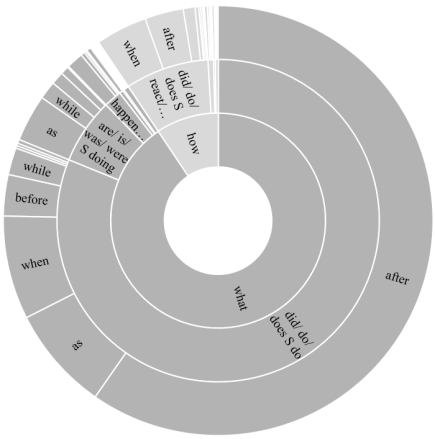

Figure 4: NeXT-QA RTS question structure. S signifies subject of the main clause.

B. SeViLA’s performance on RTS

This section presents the details concerning the data predicted by the model.

B.1. Matching in Absolute Numbers and Percentage

Target answers do not have underdetermination with little data irrelevant from the aspect point of view.

Table 4
The amount of correct and incorrect predictions of SeViLA on RTS

Matching	T	A	U	I
correct	425	758	0	12
incorrect	356	504	1	4

Table 5
The percentage of correct and incorrect predictions of SeViLA on RTS

Matching	T	A	U	I
correct	54.42	60.06	0	75
incorrect	45.48	39.94	100	25

B.2. Predicted vs. Target Answers

The examination of the most frequently predicted and target answers reveals a significant number of matches, predominantly characterised by atelic labels.

Table 6
Top 10 predicted answers

Answer	Telicity Label	Amount
smile	A	21
walk away	T	18
walks away	T	17
look around	A	17
look at camera	A	16
stand up	T	15
laugh	A	14
clap	A	11
turn around	T	10
smiling	A	8

Table 7
Top 10 target answers

Answer	Telicity Label	Amount
walk away	T	27
stand up	T	19
laugh	A	17
smile	A	17
dance	A	15
look at camera	A	15
turn around	T	12
clap	A	9
look around	T	9
smiling	A	9