Towards Semantic Representations with a Temporal Dimension*

Johanna Björklund
Dept. of Computing Science
Umeå University
johanna@cs.umu.se

Frank Drewes
Dept. of Computing Science
Umeå University
drewes@cs.umu.se

Iris Mollevik
Codemill AB
Umeå, Sweden
iris@codemill.se

Abstract

We outline the initial ideas for a representational framework for capturing temporal aspects in semantic parsing of multimodal data. As a starting point, we take the Abstract Meaning Representations of Banarescu et al. and propose a way of extending them to cover sequential progressions of events. The first modality to be considered is text, but the long-term goal is to also incorporate information from visual and audio modalities, as well as contextual information.

1 Introduction

Semantic parsing consists in translating input media into a structured representation of its meaning. Depending on the domain, the notion of 'meaning' can be more or less well-defined, but the objective is typically to capture an understanding of the input in a format that is helpful for downstream processing. Traditionally, the focus has been on textual input, but semantic parsing can also be applied to other modalities such as images, video, and sound, or, as in the multimodal case, a combination thereof. As for the output representation, previous work has evaluated various types of logics (Cai and Yates, 2013), query languages (Yu et al., 2018), vector-based embeddings (Palangi et al., 2016), graph languages (Flanigan et al., 2014), and lambda calculus (Wong and Mooney, 2007) as formal carriers of meaning.

In this abstract, we report on recently initiated work whose ultimate goal it is to develop a framework for semantic parsing of video into graph-based representations. Our work, which is a collaboration between Umeå University and the IT company Codemill AB, is motivated by the ever increasing use of video for online information and entertainment services, process control, and teleconferencing. Efficient semantic parsing for video opens for automation of countless tasks, for example, compliance checking and knowledge extraction, to name only two. On the output side, our interest is in representations based on formal graph languages, due to their close link to automata theory and suitability for algorithmic manipulation. As our starting point we take the Abstract Meaning Representation (AMR) by Banarescu et al. (2013) due to the amount and quality of publicly available language resources.

One of the characteristic features of video compared to written text is the prominent temporal aspect. In AMR graphs (AMRs, for short), nodes represent concepts and objects, and edges represent relations. However, in a video what is to be represented changes as time progresses. Hence, single AMRs cannot capture what happens in a video at a useful level of detail. This challenge is present already in the case of text. By design, AMRs capture the meaning of individual sentences. Therefore, AMRs typically make a statement about the world as seen from a particular point in time. Take for example the sentence "The woman starts out on a bench, but later she jumps in the air, and finally she lands on the ground". In the sentence, the woman is both on the bench, in the air, and on the ground, but the temporal adverbs place these events in time relative to a fixed present. Sentences such as "She is born, she lives a full life, and she dies content", which simultaneously sees the the world from three points in time, are rare exceptions, and are thus not really accounted for in AMR. In video data, however, this type of situation rather is the rule: There is not a single present, but a stream of events, and the length of sequences of events is at a completely different scale. Moreover, in AMRs, the sequence of world configurations that result from a

*This work was supported by the Wallenberg AI, Autonomous Systems and Software Program. The authors have contributed equally and are listed in alphabetical order.

sequence of events can only be extracted through a logical inference that requires ontological information. For practical reasons, it is desirable if the sequence of world configurations can, at least to a some degree, be derived more through syntactic means. Our goal is to develop a representation that makes this possible. Desiderata for the sought representation are therefore that it can (i) accommodate arbitrarily long sequences of events in a clear and compact way, (ii) describe the world as seen from an unbounded number of points in time, all equally real, and (iii) allow for easy extraction of successive world configurations.

The research presented here is only in its infancy, and we are grateful for any comments received that may help direct our efforts. We are also open for new collaborations and encourage readers interested in the topic to reach out to the authors.

2 Related Work

The first efforts to apply semantic analysis in the automation of video processing were made at the end of the last century. An example is the work by Nack and Parkes (1997), an attempt to automatically compose humorous video clips based on a library of existing footage. Since then, a central line of work has been activity recognition. The problem is addressed by Xu et al. (2005) who propose a hierarchical approach, intended to simplify the analysis by separating different levels of granularity. The authors use a framework based on Hidden Markov Models to recognise activities in sports videos, for example, a basketball shot divided over the temporal sequence of *lay up*, *shot* and *offence*. Whereas Xu et al. (2005) use annotated samples to train their system, Sener et al. (2015) present a framework for unsupervised semantic parsing of videos which identifies so-called semantic steps in the video from both video and audio data. The framework first identifies salient words and objects in the video, and then clusters these into activities. The output is a temporal sequence of labelled activity types.

Turning to graph-based approaches, Yadav and Curry (2019) represent a video stream by a stream of graphs. An ensemble of deep learning models is used to detect high-level semantic concepts from the video. Objects are identified by performing object and attribute detection on the video frames using a pipeline of deep neural networks. Each object is represented by a node in the graph; edges represent the spatial and temporal relationships between objects, calculated with spatial and temporal calculus. For each video frame, a timestamped graph snapshot is constructed. The authors propose an aggregated view of the graph stream in which the graph stream for a given time interval is represented as a single graph, as well as a method to generate such aggregated graphs. The aggregated view contains each unique object node from the time interval. Edges between nodes contain an array of different timestamped values, one for each time step (i.e. frame); for example, the distance between two cars at different time steps.

Aakur et al. (2019) use a Markov Chain Monte Carlo process where the required proposal functions are based on the ConceptNet knowledge base (Speer et al., 2017). The authors work under the hypothesis that the inclusion of common-sense knowledge can lessen the needs for training data and help reveal complex semantic relationships.

Other work on representing video content in graph form includes (Charhad and Quénot, 2004; Xin Feng et al., 2017). Instead of explicitly representing the video as a graph, one can also convert natural language queries into semantic graphs and match those against video content; this has been done by Lin et al. (2014) and Chen et al. (2019) using different techniques.

3 Representation of Temporal Semantics

Graph-based representations are attractive in semantic parsing due to their expressiveness and transparency. In line with the previously outlined desiderata, they have the potential to represent concepts and relations in a readily accessible form, while abstracting from irrelevant detail. A central question of the present work is how a temporal dimension can be added to representations such as AMR. Similar to the work of Aakur et al. (2019), this can be seen as an extension of the work by Charhad and Quénot (2004), Xin Feng et al. (2017), and Yadav and Curry (2019), which focuses on objects and their relations, to more complex semantic content including actions. Efforts in this direction may also help overcome a more fundamental limitation of AMR, namely its inability to satisfactorily capture longer pieces of text. AMR was originally designed to represent the meaning of individual sentences. AMR quickly reaches the limits of its expressiveness when made to cover entire paragraphs, sections, or chapters. It is frequently

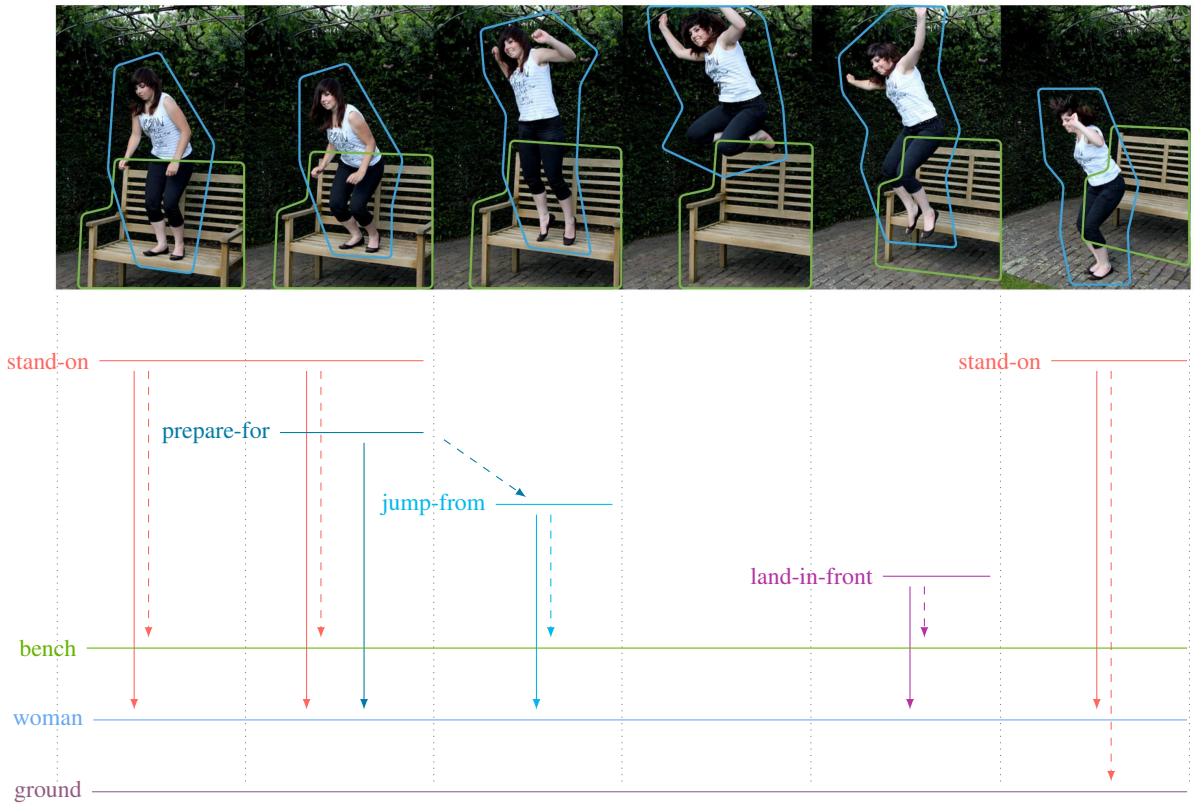

Figure 1: A graph-based temporal representation expressing that "The woman stands on the bench and prepares to jump from it, to later land on the ground in front of the bench". The image is "Sequence Jump" by Rick Beumers, licensed under CC BY-NC 2.0.

the case that these longer texts are intertwined in a way that, at least superficially, resembles the way in which the scene of a video clip develops. It therefore likely that a temporal extension would also benefit semantic parsing that focuses exclusively on text.

We propose to represent successive world configurations as ordinary AMRs, and interconnect these AMRs in a purposeful way. Nodes that represent the same entity (in separate AMRs) are identified. We model time as discrete time steps arranged in a partial order. Furthermore, we propose to assign to each node and edge in those AMRs an explicit interval of validity. Thus, this interval defines the lifespan of the node or edge. The idea is illustrated in Figure 1, where key events in an image sequence are represented as interconnected AMRs in which all objects, both entities and relations, have a demarcated existence in time. The approach can be applied to written text by equating time intervals with other sequentially arranged units, for example, sentences or paragraphs. This adds some constraints on the well-formedness of our graph; for example, the lifespan of an edge must be contained in the intersection of the intervals of validity of the nodes it connects. Thus, any framework eventually developed to support our representation must obey these constraints. A central challenge will be to decide what objects to represent, and how to understand their duration in time, not least when dealing with text.

4 Next Steps

As a first step, we formulate a framework for representing the temporal evolution of literary text, and then proceed to extend it to semantic parsing of video content. The transition to this richer domain is helped by the presence of subtitles or transcriptions of spoken dialogue. See Figure 2 for an idea of how the AMR derived from the text might act as a backbone into which information from other modalities can be fused. The long-term goal is to also incorporate information from the visual and audio modalities, as well as contextual information provided by knowledge bases such as ConceptNet (Aakur et al., 2019). In this work, we

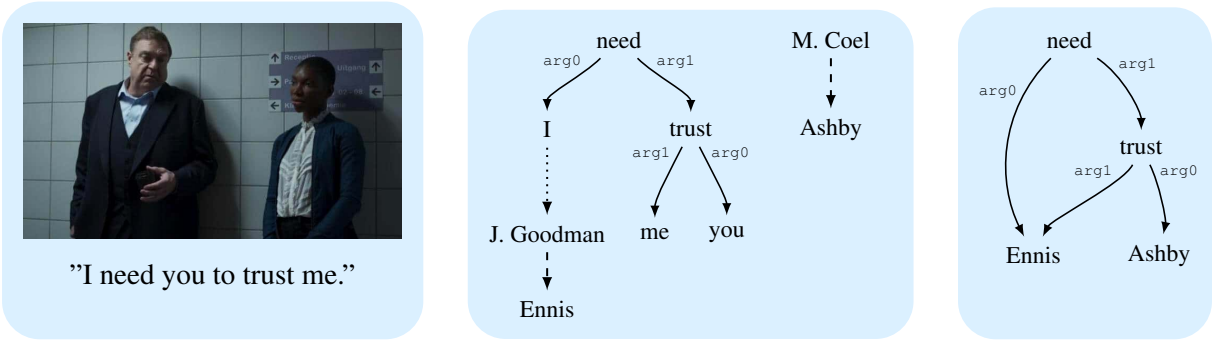

Figure 2: Speech recognition, face detection, and speaker identification are applied to the input video (left) and combined with IMDB metadata to derive a set of metadata fragments (middle), which are fused to a graph-based semantic representation (right).

use the publicly available corpus[1] of AMRs for the novella "The Little Prince" (de Saint-Exupéry, 1943). The novella also has the advantage of having been filmatised several times, which is helpful in the multimodal part of the project.

On the practical side, another goal of this project is to integrate new semantic parsing techniques into the media analysis software and workflow automation of the industrial partner Codemill AB. The aim is to derive semantic graphs that provide information for, e.g., autonomous trading of digital resources, protection against compliance violation, generating content recommendations for viewers, and automatically compiling trailers for different regions. By evaluating the methods in a real-life environment, we expect to gain insights that further the development of our representational framework, beyond what can be accomplished through purely academic research.

Acknowledgment We thank the reviewers for their helpful comments.

References

Sathyanarayanan N. Aakur, Fillipe D. M. de Souza, and Sudeep Sarkar. 2019. Going deeper with semantics: Video activity interpretation using semantic contextualization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 190–199.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proc. 7th Linguistic Annotation Workshop, an ACL 2013 Workshop*.

[1] `https://amr.isi.edu/download.html`, accessed on 2020-09-04.

Qingqing Cai and Alexander Yates. 2013. Semantic parsing Freebase: Towards open-domain semantic parsing. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 328–338, Atlanta, Georgia, USA. Association for Computational Linguistics.

Mbarek Charhad and Georges Quénot. 2004. Semantic Video Content Indexing and Retrieval using Conceptual Graphs. In *ICTTA*, Damascus, Syria.

Yuting Chen, Joseph Wang, Yannan Bai, Gregory D. Castañón, and Venkatesh Saligrama. 2019. Probabilistic semantic retrieval for surveillance videos with activity graphs. *IEEE Transactions on Multimedia*, 21(3):704–716.

Jeffrey Flanigan, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436.

Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Frank Nack and Alan Parkes. 1997. The application of video semantics and theme representation in automated video editing. *Multimedia tools and applications*, 4(1):57–83.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):694–707.

Antoine de Saint-Exupéry. 1943. *The Little Prince*. Harcourt. Translated to English by Katherine

Woods. Originally published in 1943 as *Le Petit Prince* by Gallimard.

Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4480–4488.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI 31*.

Yuk Wah Wong and Raymond Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 960–967.

Xin Feng, Yuanyi Xue, and Yao Wang. 2017. Video object graph: A novel semantic level representation for videos. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 680–685.

Gu Xu, Yu-Fei Ma, Hong-Jiang Zhang, and Shi-Qiang Yang. 2005. An HMM-based framework for video semantic analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(11):1422–1433.

Piyush Yadav and Edward Curry. 2019. VEKG: Video event knowledge graph to represent video streams for complex event pattern matching. In *2019 First International Conference on Graph Computing (GC)*, pages 13–20.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.