Representations of Time in Symbol Grounding Systems

Frank Förster and **Chrystopher L. Nehaniv**
Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire
College Lane, Hatfield, AL10 9AB, United Kingdom

Abstract

This paper gives a short overview of time representations in current symbol grounding architectures. Furthermore we report on a recently developed embodied language acquisition system that acquires object words from a linguistically unconstrained human-robot dialogue. Conceptual issues in future development of the system towards the acquisition of action words will be discussed briefly.

Introduction

Symbol grounding systems recent years have focused primarily on the grounding of object and action words. Object words usually denote physical entities that are relatively persistent in time and whose meaning typically does not include any reference to time - a red ball will typically stay red and round for a potentially infinite time-period and its affordances with regard to an agent will typically remain unchanged. The same cannot generally be said of action words which denote events that are characterized through changes in the spatial location of an agent and/or object, changes in its physical configuration or changes in the spatial and/or physical relationship between two or more agents and objects. Symbol grounding systems that cover action words, i.e. that associate the linguistic label with sensorimotor data from the agent, must therefore represent time in some way. Some systems use logic based formalisms in order to model events and actions (Siskind 2001; Steels and Baillie 2003), some are based on recurrent neural networks (Sugita and Tani 2005) and some employ both approaches on different levels (Dominey and Boucher 2005). A third kind of symbol grounding architecture which is currently developed by Saunders et al is inspired by memory-based natural language processing in order to acquire and ground object words taken from a linguistically unconstrained human-robot dialogue (Saunders, Nehaniv, and Lyon 2010).

Time in Logic Based Systems

There are several common means in temporal logics to state factual assertions about states of affairs whose truth is limited to certain points or intervals in time. Point-based temporal representations mark explicitly the beginning and end of an event (e.g. *touch(t1,t2)*) on a linear time scale as opposed to time-interval based representations like *touch(T1)*, where *T1* is a time-interval. Modal tense operators mark temporal properties by specific operators like *PAST(ball(red))* and were used in the past in order to mark tense in natural language but have severe limitations by being rather coarse grained with regard to temporal resolution (Allen 1991).
Recent symbol grounding architectures like (Siskind 2001; Dominey and Boucher 2005) use event logic, Steels and Baillie (Steels and Baillie 2003) use the temporal interval calculus to represent time. The former is an example for a point-based temporal representation, whereby Siskind introduced the notion of *spanning intervals* and developed a inference procedure. Spanning intervals were introduced in order overcome the computational complexity caused through at least quadratically many subintervals of an interval associated with a *liquid event*. Roughly spoken events are liquid if they hold for any subinterval of i, if they hold for an interval i.
The temporal interval calculus is principally based on time-intervals but also allows the representation of points in time.

Time in Connectionist Systems

In connectionist models time can be either represented explicitly or implicitly. The explicit case can be regarded as an instantiation of the time-space-metaphor: data that is originally ordered in time is mapped onto a spatial order (sequence) and presented simultaneously to the neural network. Recurrent connections amongst neurons are not necessarily needed in this case. A severe drawback of this approach is that the length of the time series is limited to the number of neurons in the input layer.
In the implicit case recurrent connections are made between several neurons of a given network depending on the architecture. Instead of mapping the temporal onto a spatial order, the data is presented to the network in its original temporal order. Time is in this case implicit in the way that the data is processed by the network (see (Elman 1990)). Recent connectionist symbol grounding architectures like the PBRNN introduced by Sugita and Tani (Sugita and Tani 2005) are examples of networks whose time representation is implicit.

Time in Memory-based Systems

Memory-based methods differ from the methods discussed above in that the data is not replaced by an abstract model

after a given learning or model construction phase. To speak of time representation in this case might be slightly misleading as there is no model and therefore no representation beyond the format of the recorded data. Saunders et al (Saunders, Nehaniv, and Lyon 2010) developed a memory-based system for symbol grounding which is part of a language acquisition system. This system extracts utterances from human language which originate from human-robot dialogues. These dialogues take place across several sessions with increasingly correct linguistic activity on the part of the robot. After each session the system associates the extracted words with sensorimotor data. In the consecutive session the robot engages in an utterance if its sensorimotor state is *similar enough* to a former state in which the human dialogue partner made this very utterance. Whether a sensorimotor state is *similar enough* to another is detected by a k-nearest-neighbor (kNN) algorithm, which operates on the readings of the sensorimotor stream with weighted attributes (see (Saunders, Nehaniv, and Lyon 2010) for details). Consequently the utterances can be regarded as labels of sensorimotor patterns. Algorithmically the resulting situation can be regarded as a pattern matching problem.
So far the system focused on the acquisition of utterances that refer to objects. Timing was only important in the sense of a correct alignment of the stream of linguistic with the stream of sensorimotor data. The system is currently being extended towards the acquisition of action words. Consequently ways have to be found to detect temporal patterns in multivariate time-series instead of detecting *static* patterns in vectors of data that originate from discrete time points.

Challenges

Other scientific fields which perform analyses on multivariate time-series advanced distance measures to be used with kNN are discussed by (Keogh and Ratanamahatana 2005; Yang and Shahabi 2007). Although the fact that the labels are words in our scenario create a specific problem that seems to be unique. Words have at least two properties which are problematic from a pattern matching perspective.

- Words are vaguely defined or not defined at all. Looking at one word alone is most probably not enough as they are characterized through family resemblances (see (Wittgenstein 1958)). For pattern matching purposes this means in an extreme case that all exemplars do not share a single common property despite having the the same label. On the positive side we do know from the concept of family resemblance that at some point at least one of the already existing and labeled exemplars must have a common property with an new exemplar that should be labeled identically. "Some point" here means, that the latter might not be the case when there are only few exemplars in the repository of already labeled sequences.
- Word labels are not unique labels. The same sequence of sensorimotor states can be labeled with many words. Imagine the situation of a red toy car, which is pushed and rolls from left to right stalling on the right side. The first part of the sequence could be labeled with *push*, the last part with *stop* or *roll out*. The middle part of the sequence could be labeled with *move* or *roll* but as well with *drive*. *Drive* and *roll* are only distinguished by intentionality (driver vs. no driver) that might not show up in the data at all.

In case of action or event words like "push", "roll", or "stop" there is an additional variance in terms of when they stop or start. "Push" has a clear-cut beginning but a rather fuzzy end. To denote an action with "push", it has to be successful. Being successful in this context means that the pushee has to actually move for some time. The problem is that we cannot say precisely for how long the object has to move in order for a push to be a push.
The variance of events in terms of their temporal length might be tackled by dynamic time warping. The latter does not assume that two multivariate time series have to be of the same length. Also the the phase of the pattern can be shifted and the similarity will still be detectable by warping the time-axis (see (Sakoe and Chiba 1978)).

Acknowledgments The work described in this paper was conducted within the EU Integrated Project ITALK ("Integration and Transfer of Action and Language in Robots") funded by the European Commission under contract number FP-7-214668.

References

Allen, J. F. 1991. Time and time again: The many ways to represent time. *International Journal of Intelligent Systems* 6(4):341–355.

Dominey, P., and Boucher, J. 2005. Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence* 167:31–61.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14(2):179–211.

Keogh, E., and Ratanamahatana, C. A. 2005. Exact indexing of dynamic time warping. *Knowledge and Information Systems* 7(3):358–386.

Sakoe, H., and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing* 26(1):43–49.

Saunders, J.; Nehaniv, C. L.; and Lyon, C. 2010. Robot learning of object semantics from the unrestricted speech of a human tutor. *submitted*.

Siskind, J. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research* 15:31–90.

Steels, L., and Baillie, J. 2003. Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems* 43:163–173.

Sugita, Y., and Tani, J. 2005. Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior* 13(1):33–52.

Wittgenstein, L. 1958. *Philosophical Investigations*. Oxford: Blackwell.

Yang, K., and Shahabi, C. 2007. An efficient k nearest neighbor search for multivariate time series. *Information and Computation* 205(1):65–98.