Incremental Unit Networks for Multimodal, Fine-grained Information State Representation

Casey Kennington
Department of Computer Science
Boise State University
caseykennington
@boisestate.edu

David Schlangen
Department of Lingusitics
University of Potsdam
david.schlangen
@uni-potsdam.de

Abstract

We offer a sketch of a fine-grained information state annotation scheme that follows directly from the Incremental Unit abstract model of dialogue processing when used within a multimodal, co-located, interactive setting. We explain the Incremental Unit model and give an example application using the Localized Narratives dataset, then offer avenues for future research.

1 Introduction

Human experience is profoundly multimodal. As people explore the world they are organizing perception, action, and thought in a complex social environment (Smith and Gasser, 2005). Tied directly to this multimodal experience is human language, primarily spoken language (Fillmore, 1981), and a growing body of literature across several disciplines make a strong case that language learning and language meaning is grounded in rich multimodal (even embodied), interactive, and enactive experience (Pulvermüller, 1999; Barsalou, 2008; Smith and Samuelson, 2009; Di Paolo et al., 2018; Bisk et al., 2020). Despite this, current state-of-the-art language models such as BERT (Devlin et al., 2018) are trained only using static text, and while it is clear that such models are powerful and useful for many tasks, they are clearly missing important multimodal semantic knowledge (Rogers et al., 2020; Bender and Koller, 2020).[1] We argue that what is needed is a semantic model that is learned not only from text, but has knowledge of multiple modalities and that the model operates in a setting similar to how language is acquired for humans: multimodal, co-located, interactive spoken dialogue:

multimodality: A model of semantic meaning of language must ground into not just vision, but other modalities such as taste, touch, smell, proprioperception, and even affect. This is as much a modeling challenge as an engineering challenge, because each modality requires sensor hardware (e.g., cameras for vision) and methods for fusing the sensor information from different modalities.
co-location: Multimodal systems have multiple sensors that sense things like objects, events, and the interlocutor who has knowledge about the environment, language used to denote objects, and uses cues such as gaze and gestures in communication.
spoken interaction: Semantic meaning is learned and used in coordination with members of a particular language community (Clark, 1996) and spoken interaction is the setting where children learn language. Moreover, spoken language differs dramatically from written text in that spoken language contains communicative artifacts such as hesitations, false starts, repetitions, repairs, and coordination of turn-taking. Furthermore, people produce and understand language sequentially, not as complete and fully grammatical units (Tanenhaus and Spivey-Knowlton, 1995).

Taken together, these requirements imply technical and modeling challenges. Technical challenges include using multiple sensors and articulators, fusing their information streams, temporally aligning input and output. Modeling challenges include binding information from the sensors, learning meaningful patterns in a noisy setting, and representing the states of the sensors and unfolding interaction.

In this paper, we don't formulate a semantic model, but focus rather on a representation with a fine-grained information state update approach using the Incremental Unit abstract model of spoken dialogue. We explain the Incremental Unit model in the next section, including how multi-

[1] Though there have been recent efforts to augment language models with some modalities such as vision, e.g., Lu et al. (2019).

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 89–94

modal how information is represented, then offer a simple scheme for using Incremental Units as a basis for developing multimodal semantic models.

2 The Incremental Unit Framework

The *Incremental Unit* (IU) framework (Schlangen and Skantze, 2011) is an abstract, conceptual approach for incremental processing for spoken dialogue. The IU framework consists of a network of processing *modules*, each of which play a different role in an unfolding dialogue, all of which work together to create the fine-grained information state. Modules take input data on their *left buffers*, process the input, then produce output on their *right buffers*. A critical part of the IU framework is how the data are packaged and processed. The data are packaged as the payload of *incremental units* (IUs) which are passed between modules–each IU holds a discrete amount of information.

Another critical part of the framework is that the IUs themselves are interconnected via *same level links* (SLL)–allowing the linking of IUs as a growing sequence–and *grounded-in links* (GRIN) which allow that sequence to convey what IUs directly affect another IU. Ideally, IUs (e.g., produced from a sensor or processing module) can be guaranteed to be correct, but often an IU that has been outputted to the next module needs to be updated in light of new information. To make this possible, the framework makes use of three operations: IUs can be *added* to the IU network, but can be later *revoked*, and also *committed* when a module can guarantee that an added IU will not be revoked.

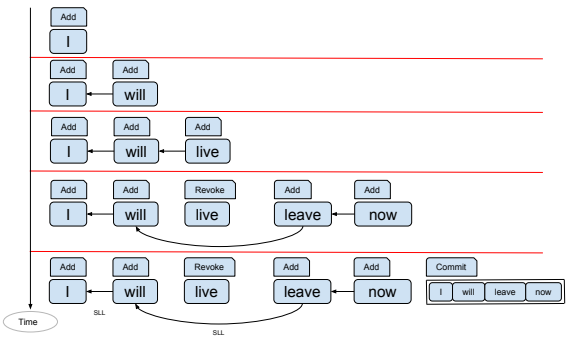

Figure 1: Example of SLL, and Add, Revoke and Commit operation for an incremental speech recognizer.

Figure 1 shows an example of how a speech recognition module would process incrementally, typically word-by-word. It takes a continuous audio signal as input from a microphone and produces discrete word IUs as output. As the utterance *I will leave now* is uttered, the speech recognizer outputs words as they are recognized at the word level and adds them to the IU network. The recognizer mis-recognized the word *live*, but in light of new information from the unfolding utterance, revoked *live* and replaced it with *leave*. Horizontal arrows show SLLs; i.e., how the IUs are related to each other temporally, and at the end of the utterance when the recognizer knows it will no longer revoke, it marks all of the IUs as committed. IUalso contain information about their creation time.

It's important to distinguish at this point the networked IU *modules* or processors that pass IUs to each other and the network of IUs themselves. For example, a speech recognizer might pass its transcribed speech as IUs with payloads of word strings to a part-of-speech module that produces a part-of-speech for each word as payloads of part-of-speech strings, which are then the input of a language understanding component that operates on both the words and parts-of-speech to produce some kind of semantic abstraction of the unfolding utterance. Thus the three processors–speech recognizer, part-of-speech tagger, and language understanding–are separate modules, but each use the *add, revoke, commit* operations to alter the shared network of IUs. The IU framework, including the operations, can be used as a fine-grained model of the dynamics of the creation of the information state of an agent in a situated interaction, comprising both its world model and its discourse model, and the interaction between them.

Multimodal Example Following Kennington et al. (2014), Figure 2 shows an example of modules and IUs created by a multimodal system co-located with a human interlocutor. For this example, the system is tasked with learning about objects. In this specific turn of the interaction, the interlocutor utters *this is my phone* accompanied by a display of the phone and a deictic pointing gesture. The system has two sensors, a camera and a microphone. The microphone feeds continuous audio to the automatic speech recognizer (ASR), which transcribes the utterance into word IUs. Those are outputted to a part-of-speech (POS) tagger that produces part-of-speech IUs. Those in turn are outputted to the semantic parser (SEM) which produces a semantic abstraction over the utterance; the semantic parser uses both words and parts-of-speech to produce the under-specified semantic parse IUs. Those are given to a natural

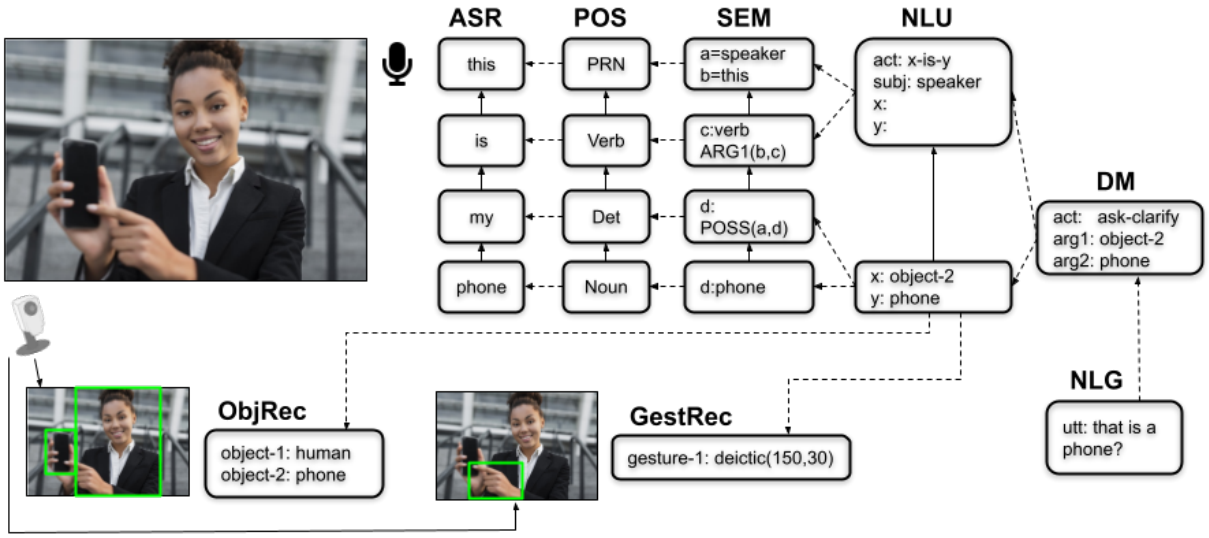

Figure 2: Example of a system made up of two modalities (i.e., audio and vision), camera and microphone sensors, and processing modules. An interlocutor says *this is my phone* accompanied by a deictic gesture to the phone; the modules process the scene and audio; the DM (dialogue manager) makes a decision to ask a clarification question which is rendered by the NLG as *is that a phone?*. The modules create the IUs, which are connected to each other via same-level links (solid lines) and grounded in links (dashed lines), the latter denote the IUs that played a role in that IU's creation. For example, the bottom IU for NLU needed information from IUs created by the ObjRec, GestRec, and SEM modules. The full network constitutes a multimodal meaning representation.

language understanding (NLU) module that produces a semantic frame (that is more closely tied to the particular task of learning new words), and the dialogue manager (DM) makes a decision about the action to take next; in this case it decides to ask a question to the user about the denoted object and the associated word, then the natural language generation (NLG) formulates the utterance that is uttered through a speaker using a speech synthesizer to the interlocutor.

Prior Work As a theoretical model, the IU framework formed the basis for a model of temporally aligning different sensor modalities; Kennington et al. (2017a) showed that timestamp information in the IUs can be used to inform modules to add IUs to the IU network at the same time, thereby giving downstream modules information about an event that may have happened, even if the sensors produced processing delays. Buß and Schlangen (2011) leveraged the IU operations for an incremental dialogue manager that could make self corrections (e.g., if the system began an utterance, but a revoke meant that the utterance should change, the system would self-correct), and Lison and Kennington (2017) used the IU operations to inform a neural conversation model. The IU framework has also been the inspiration for several spoken dialogue system architectures, and several implementations based on the IU framework have been developed. InproTK (Baumann and Schlangen, 2012) is the most commonly used (written in Java), and was extended to incorporate modalities beyond just speech (Kennington et al., 2017b). More recently, ReTiCo (Michael and Möller, 2019) was developed (written in Python) and extended to incorporate multiple modalities, evaluated in a multimodal robotic system (Kennington et al., 2020).

Using a network (or a graph) to represent meaning has received recent attention, yet has a long history. Koller et al. (2019) provides an overview of several formalisms, including Abstract Meaning Representation (Banarescu et al., 2013), a particular representation that has seen adoption in the community. However, these graph-based semantic representations are focused only on representing sentences, not multimodal information, and does abstract away from the dynamics of creating the network.

3 The IU Framework for Fine-grained Information State Representation

In this section, we sketch a scheme for the IU network as a representation of a fine-grained information state. The scheme follows the IU approach to processing live speech; all annotations are packaged as IUs with links between them, all *add* op-

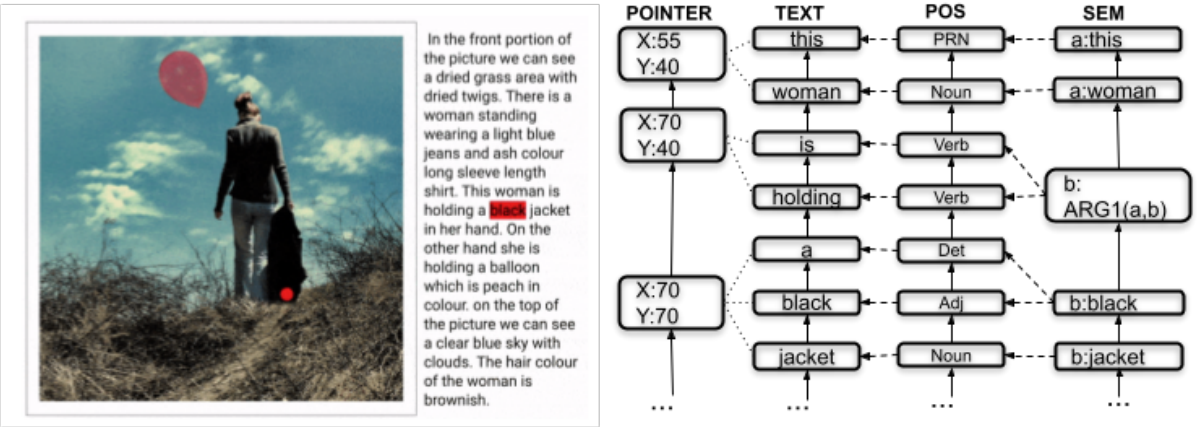

Figure 3: Example of Pointer, Word, POS, and SEM IU annotations for a sample from the Localized Narrative dataset. Solid lines denote SLLs, dashed denote GRINs, and the dotted lines denote an alignment between two modalities. Image taken from https://google.github.io/localized-narratives/.

erations are accounted for (and *revoke* operations under live annotation conditions), each operation is timestamped, and the creation time of each IU is timestamped. We don't specify how the modalities or modules interact with each other, the goal here is to focus on the information state.

We give an example in Figure 3 using a sample from the Localized Narratives dataset (Pont-Tuset et al., 2020). The dataset consists of images described by annotators. Descriptions have speech and mouse pointer modalities that are later temporally aligned. Speech is automatically transcribed as the annotators speak, but annotators are tasked with hand-transcribing their descriptions after they are complete. The dataset on its own has multimodal annotations, though it's unclear how they would work in a live interaction with a system.

The IU network annotation in Figure 3 shows locations of mouse pointer (x,y coordinates), words, and added part-of-speech tags and semantic abstraction similar to that in Figure 2. The SLL and GRIN links are also present, and additional links between the speech and pointer modalities are depicted. What is not depicted in the figure are the *add* and *revoke* operations that enable the network to grow as an interaction unfolds in real time, though it is obvious that all IUs in the figure were created through an *add* operation. In the case where a perfect transcription exists, only *add* operations are necessary, but a live interaction would require the ability to *revoke* erroneous words then *add* correct ones in real time, in alignment with the movements of the mouse pointer. Timestamp information is not present in the figure; time generally flows downward as IUs are added to the network.

The scheme can be applied during the data collection process. This requires some up-front effort to setup each individual module to operate incrementally. For the Localized Narratives dataset, incremental text can come from ASR or typed text, and the other annotations from respective modules. Annotated data can be represented in any format, e.g., JSON. This scheme highlights the importance of annotating data that is representing a fine-grained information state collected in a multimodal, co-located, and spoken interactive task. Such a representation is potentially useful for a formal representation of situated conversation and embodiment.

4 Conclusion and Future Work

In this paper, we outlined an IU network-based approach to representing multimodal states within the requirements of multimodality, co-location, and interactive speech. Implicit in this representation is the requirement that the system is modular, though it is potentially possible to represent the IU network in an end-to-end neural architecture. The modalities explored here were only a minimal example of what the network could potentially handle– added modalities enrich the semantic representation. For example, we have used the IU framework to represent audio, visual, and internal robot state modalities in prior work (Kennington et al., 2020). We leave formalizing semantic operations, such as compositionality, meaning derived from handling uncertainty or requests for clarification, and global decoding strategies in the IU network semantic representation for future work.

Acknowledgements We appreciate the feedback from the anonymous reviewers.

References

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 *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Lawrence W Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology*, (59):617–645.

Timo Baumann and David Schlangen. 2012. The InproTK 2012 release. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 29–32.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Association for Computational Linguistics*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. *arXiv*.

Okko Buß and David Schlangen. 2011. DIUM – An Incremental Dialogue Manager That Can Produce Self-Corrections. In *Proceedings of semdial 2011 (Los Angelogue)*, Proceedings of semdial 2011 (Los Angelogue).

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Ezequiel A Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. 2018. *Linguistic bodies: The continuity between life and language*. Mit Press.

Charles J. Fillmore. 1981. Pragmatics and the description of discourse. *Radical pragmatics*, pages 143–166.

Casey Kennington, Ting Han, and David Schlangen. 2017a. Temporal Alignment Using the Incremental Unit Framework. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 297–301, New York, NY, USA. ACM.

Casey Kennington, Ting Han, and David Schlangen. 2017b. Temporal alignment using the incremental unit framework. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 297–301, New York, NY, USA. Association for Computing Machinery.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. Situated incremental natural language understanding using a multimodal, linguistically-driven update model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1803–1812, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. 2020. rrSDS: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 132–135, 1st virtual meeting. Association for Computational Linguistics.

Alexander Koller, Stephan Oepen, and Weiwei Sun. 2019. Graph-based meaning representations: Design and processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–11, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Casey Kennington. 2017. Incremental Processing for a Neural Conversational Model. In *Proceedings of SemDial*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.

Thilo Michael and Sebastian Möller. 2019. ReTiCo: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *Tagungsband der 30. Konferenz Elektronische Sprachsignalverarbeitung 2019*, ESSV, pages 134–140, Dresden. TUDpress, Dresden.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Proceedings of ECCV*.

Friedemann Pulvermüller. 1999. Words in the brain's language.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *arXiv*.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. In *Dialogue & Discourse*, volume 2, pages 83–111.

L B Smith and L Samuelson. 2009. Objects in Space and Mind: From Reaching to Words. In *The Spatial Foundations of Language and Cognition*.

Linda Smith and Michael Gasser. 2005. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial Life*, (11):13–29.

Michael K Tanenhaus and Michael J Spivey-Knowlton. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632.