SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 186–187.
Los Angeles, California, 21–23 September 2011.

Tracking Communication and Belief in Virtual Worlds

Antonio Roque
Computer Science Department
University of California, Los Angeles
aroque@ucla.edu

Abstract

We are developing an approach to determining the gist of interactions in virtual worlds. We use algorithms to extract and combine virtual world features into various types of evidence of understanding, which are used by individuals to develop their beliefs about the world and its events.

1 Virtual World Interactions

Virtual Worlds are valuable research platforms: they provide embodied situated language use, they include persistent user profiles, and they contain lower-noise alternatives to real-world Automated Speech Recognition and Object Detection technologies. Virtual Worlds are also inherently interesting because they are used by a large number of people, many of them children. Of the 1.2 billion registered accounts in public virtual worlds, over 730 million of them belong to users under the age of 15 (KZero Corporation, 2011). This is because virtual worlds are more than standalone applications offering full 3D graphics; they now include web-browser-based 2.5D worlds, which are often marketed along with real-world toys.

However, automatically determining the gist of virtual world interactions is not trivial. Consider two case studies that highlight the difficulty of capturing the essence of an interaction in a virtual world.

First, imagine one virtual character telling another: "I have the package for you to take." It is not enough to say that the utterance contains a statement or an implied command, or even whether the utterance is the result of an adversarial negotiation. Instead, we may be interested in determining whether or not this is part of a planned illegal activity. Second, imagine one virtual character telling another: "We can try this in real life tomorrow." This may be a harmless social statement, or it may be the behavior of a sexual predator.

Such utterances occur in interactions between agents who share a rich context. To identify the nature of the interaction, we need to model the situated world context, the relational history between the virtual characters, and their shared knowledge, for example. When possible, we would like to distinguish between what the speaker believes is meant and whether the hearer and overhearers share that belief: for example, whether everyone knows what exactly is in a package being discussed.

We would like our model to be updated in real-time to integrate new activities as they occur, and to be explainable so that a human can trace the reasons for the model's conclusions. We would also like this approach to be platform-neutral, so that it can be adapted to new virtual worlds as they are developed, as well as to 2.5D web-based worlds, online games, interactive texts, or potentially even video streams.

As described in the next section, we are developing an approach in which meaningful features are extracted from an interaction in a virtual world and used to build a model an online population's beliefs and utterances. This population model can then be queried to identify the beliefs of the agents regarding the interactions that they have experienced.

2 Approach

In the first stage, low-level data perceived in the virtual world is used to extract higher-level features. For example, imagine that three characters are gathered around an in-world object,

and that one of the characters makes an utterance. The low-level data includes the relative positions of each character, the direction the characters are facing, and the identity of the character who made the utterance, for example.

To extract higher-level features, we may calculate which characters were in the "hearing" range of the utterance (assuming the utterance was made by an in-world chat with a range), whether the utterance was addressed to anyone, how the hearers reacted (by replying in a way that confirmed their understanding, or by a general acknowledgment, for example), the history of the in-world object (i.e. who created it, which characters interacted with it or referred to it, etc.) and what the utterance tells us about the relationships between the characters (is one of them an expert or more senior, for example.)

Following research in dialogue grounding (Clark and Marshall, 1981) we recognize that humans use *copresence heuristics*, or indications of information that is mutually available to all relevant individuals, to track and reason about the beliefs of other individuals Copresence heuristics are derived from low-level sensor data as described above — in a computer, they include dialogue features identified through natural language processing, and physical copresence features derived from vision and positional information. One innovation of this project is the use of copresence heuristics on a continual flow of captured network traffic to automatically build an explainable representation of the set of mutual beliefs among the individuals in a population. We investigate the different types of features available, such as visual and positional data, sounds, voice chat, and text chat.

Individuals transmit sensory information to each other while communicating and coordinating interactions in a virtual world. Our algorithms are meant to interpret this information in the same way that humans do: by integrating sensory information into evidence of understanding that represent mutual belief. The features extracted from the virtual world are combined into beliefs organized by agent. The population model contains the set of agents seen, along with that agent's beliefs, stored along with the evidence for those beliefs. That population model may then be queried in an interactive interface.

3 Related Work

Leuski and Lavrenko (2006) address one aspect of the problem by identifying an in-game action in a virtual world. Related research in activity recognition, such as by Chodhury et al. (2008), approaches the problem as one of processing and selecting features from sensors, with a classification module that uses the features to identify the activity of interest. However, feature selection is challenging: automatic approaches limit explainability and require large amounts of training data, and manual approaches may not generalize. We avoid these problems by using features derived from psychological models of human communication. Similarly, Orkin and Roy (2007) describe a statistically learned model of context that they called common ground, but that consisted only of a plan representation rather than beliefs, and which was learned offline.

Acknowledgments

This work has been sponsored by a grant issued by the IC Postdoctoral Research Fellowship program.

References

Choudhury T, Borriello G, Consolvo S, et al., (2008) "The Mobile Sensing Platform: An Embedded Activity Recognition System", IEEE Pervasive Computing, 7(2):2-41.

Clark H, Marshall C, "Definite reference and mutual knowledge", In: Elements of Discourse Understanding (1981), pp. 10-63, Joshi A and Webber I, eds.

KZero Corporation, (2011) "VW registered accounts for Q1 2011 reach 1.185bn," Accessed June 22, 2011, http://www.kzero.co.uk/blog/?p=4580

Leuski A and Lavrenko V, (2006) "Tracking dragon-hunters with language models." In Philip S. Yu, Vassilis Tsotras, Edward Fox, and Bing Liu, editors, Proceedings of the 15th Conference on Information and Knowledge Management (CIKM).

Orkin J and Roy D, (2008) "The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online." Journal of Game Development.