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.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 188–189.
Los Angeles, California, 21–23 September 2011.

Towards Speaker Adaptation for Dialogue Act Recognition

Congkai Sun
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094-2536
csun@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094-2536
morency@ict.usc.edu

1 Introduction

Dialogue act labels are being used to represent a higher level intention of utterances during human conversation (Stolcke et al., 2000). Automatic dialogue act recognition is still an active research topic. The conventional approach is to train one generic classifier using a large corpus of annotated utterances (Stolcke et al., 2000). One aspect that makes it so challenging is that people can express the same intentions using a very different set of spoken words. Imagine how different the vocabulary used by a native English speaker or a foreigner can be. Even more, people can have different intentions when using the exact same spoken words. These idiosyncratic differences in dialogue acts make the learning of generic classifiers extremely challenging. Luckily, in many applications such as face-to-face meetings or tele-immersion, we have access to archives of previous interactions with the same participants. From these archives, a small subset of spoken utterances can be efficiently annotated. As we will later show in our experiments, even a small number of annotated utterances can make a significant differences in the dialogue act recognition performance.

In this paper, we propose a new approach for dialogue act recognition based on reweighted domain adaptation inspired by Daume's work (2007) which effectively balance the influence of speaker specific and other speakers' data. We present a preliminary set of experiments studying the effect of speaker adaptation on dialogue act recognition in multi-party meetings using the ICSI-MRDA dataset (Shriberg, 2004). To our knowledge, this paper is the first work to analyze the effectiveness of speaker adaptation for automatic dialogue act recognition.

2 Balanced Adaptation

Different people may have different patterns during conversation, thus learning a single generic model for all people is usually not optimal in dialogue act recognition task. In this work, for each speaker, we construct a balanced speaker adapted classifier based on a simple reweighting-based domain adaptation algorithm from Daume (2007).

Model parameters are learned through the minimization of the loss function defined as the sum of log likelihood on speaker specific data and other speakers' data

$$Loss = w \sum_{n \in S} log(p(y_n|x_n)) + \sum_{m \in O} log(p(y_m|x_m)). \tag{1}$$

S is a set containing all labeled speaker-specific dialogue acts, O is a set containing all other speakers' labeled dialogue acts. w is for balancing the importance of speaker specific data versus other speaker's data. x_n and x_m are the utterances features, y_n and y_m are the dialogue act labels, $p(y_n|x_n)$ and $p(y_m|x_m)$ are defined as

$$p(y|x) = exp(\sum_i \lambda_i f_i(x, y))/Z(x). \tag{2}$$

3 Experiments

In this paper, we selected the ICSI-MRDA dataset (Shriberg, 2004) for our experiments because many of its meetings contain the same speakers, thus making it better suited for our speaker adaptation study. ICSI-MRDA consists of

Models	200	500	1000	1500	2000
Generic	76.76%				
Speaker only	64.07%	65.99%	68.51%	69.99%	71.06%
Simple speaker adaptation	76.81%	76.96%	77.00%	77.23%	77.53%
balanced speaker adaptation	**78.17%**	**78.29%**	**78.67%**	**78.74%**	**78.47%**

Table 1: Average results among all 7 speakers when train with different combinations of speaker specific data and other speakers' data and vary the amount of training data to be 200, 500, 1000, 1500 and 2000.

75 meetings, each roughly an hour long. From these 75 meetings, we selected for our experiments 7 speakers who participated in at least 10 meetings and spoke more than 4,000 dialogue acts. From the utterance transcriptions, we computed 14,653 unigram features, 158,884 bigram features and 400,025 trigram features. Following the work of Shriberg et al. (2004), we used the 5 general tags *Disruption*(14.7%), *Back Channel(*10.20%*)*, *Floor Mechanism*(12.40%), *Question*(7.20%) and *Statement*(55.46%) as labels. The total number of dialogue acts for all 7 speakers was 47,040.

All experiments were performed using hold-out testing and hold-out validation. Both validation and testing sets consisted of 1000 dialogue acts from meetings not in the training set. In our experiments, we analyzed the effect of training set size on the recognition performance. The speaker-specific data size varied from 200, 500, 1000, 1500 and 2000 dialogue acts respectively. When training our balanced adaptation algorithm described in Section 2, we validated the balance factor w using the following values: 10, 30, 50, 75 and 100. The optimal balance factor w was selected automatically during validation. The following four experiments are intended to prove the effectiveness of speaker balanced adaptation. Their respective results are listed in Table 1.

1. **Generic** represents the conventional method where a large corpus is used to train the recognizer and then tested on a new person who is not part of the training. The average accuracy over the 7 participants is 76.7%.

2. **Speaker Only** represents the approach where we train a recognizer using only one person data and test on spoken utterances from the same person. We show in Table 1 the average accuracy over our 7 participants for different size of training sets. Even with 2000 speaker-specific dialogue acts for training, the best accuracy is 71.06% which is much lower than 76.76% from the generic recognizer. Given the challenge in labeling 2000 speaker-specific annotated dialogue acts, we are looking at a different approach where we need less speaker-specific data.

3. **Simple speaker adaptation** represents the approach where the training set consists of all the generic utterances(from other participants) and a few utterances from the speaker of interest(same speaker used during testing). This approach is equivalent to keeping a balance factor w of 1 in equation (1). Results showing that for all 7 speakers, the accuracy always improve when including speaker-specific data with all other speakers' data for training.

4. **Balanced speaker adaptation** shows the results for balanced adaptation algorithm described in section 2. This algorithm shows significant improvement over all the other approaches in Table 1 even with only 200 speaker-specific dialogue acts. These results show that with even a simple adaptation algorithm we can improve the automatic dialogue act recognition.

References

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol V. Ess-dykema and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26:339-373.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *HLT-NAACL SIGDIAL Workshop*.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.