Using Language Groundings for Context-Sensitive Text Prediction

Timothy Lewis, Amy Hurst, Matthew E. Taylor, & Cynthia Matuszek
tim22@umbc.edu | amyhurst@umbc.edu | taylorm@eecs.wsu.edu | cmat@umbc.edu

Abstract

In this paper, we present the concept of using language groundings for context-sensitive text prediction using a semantically informed, context-aware language model. We show initial findings from a preliminary study investigating how users react to a communication interface driven by context-based prediction using a simple language model. We suggest that the results support further exploration using a more informed semantic model and more realistic context.

Keywords— Grounded language, context sensitive generation, predictive text

1 Introduction

Advances in natural language and world perception have led to a resurgence of work on the *language grounding problem*. Most work to date has focused on learning a model of language describing the world, then using it to understand novel language, e.g., following directions, (Artzi and Zettlemoyer, 2013; MacGlashan et al., 2015) or learning to understand commands in a space of plans or commands (Misra et al., 2014).

Generating language based on context is arguably more difficult, although the additional information provided by context makes this a promising area for natural language generation in general. There is a growing body of work on context-based generation in limited domains, such as sportscasting (Chen and Mooney, 2010), asking questions (Tellex et al., 2014), or generating spatial descriptions or narratives (Huo and Skubic, 2016; Rosenthal et al., 2016). In order to provide communication suggestions for users, it is not necessary to solve the problem of arbitrary natural language generation. Instead, the system must be able to provide predictions that support a predictive language interface, in which a user is continuously provided with a set of suggestions for possible speech.

We propose an approach, in which a joint linguistic/perceptual model is used to drive a predictive text tool, targeting augmentative and alternative communication (AAC) tools for wheelchair users with motor apraxia of speech. We propose to use the speaker's environment as *context* to make more relevant predictions.

Sensed context will be used to drive the probability of predictions and reduce ambiguity; for example, while "button" may refer to a fastener for clothing or a control for an electronic device, someone in front of an elevator is probably referring to the latter, which in turn focuses what they are likely to want to say. Instrumented wheelchairs can capture a large corpus of language paired with context to support development of a user-specific model trained before and during degradation of the ability to speak.

This paper discusses a pilot study using a preliminary language model with simulated context. Participants responded to scenarios using a prototype interface to communicate. Using results and observations from this user study, we hypothesize that context-based predictive lan-

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 48–52,
Austin, TX, November 5, 2016.

guage can improve usability of a predictive text interface and represents a promising direction for future work.

2 Approach

In grounded language acquisition, a combination of language and physical context are used to develop a language model for understanding future utterances. (Mooney, 2008) The context can be physical (depending on physical sensors, sometimes on a robot), (Fasola and Mataric, 2014) a simulation of some physical context, (Chen and Mooney, 2011) or more abstract descriptions. (Kress-Gazit and Fainekos, 2008) We propose to collect and learn from a similar set of data, with a language model targeting generation rather than understanding.

2.1 Corpus Collection

In order to learn a model of contextualized speech, it is necessary to collect both spoken language and context describing the environment when communication is occurring. We propose to perform this collection in three stages, from general to user-focused, as we build a better corpus and model.

(1) Crowdsourcing To gain a better understanding of how people may respond in different situations, Mechanical Turk will be leveraged to solicit responses from users about various scenarios. Each scenario presents a speaker with text describing a certain situation (and images when appropriate) and asked what they would say. This provides us with an initial corpus of typed responses to known scenarios. The preliminary study (see Section 3) was performed on a small-scale crowdsourced corpus.

(2) Telepresence For the second stage, we will use a telepresence robot (see Figure 1). The Beam robot provides insight into situations that may require assistance (for example, having the robot travel between floors of a building via the elevator, or delivering a package from one office to another). The Beam's existing video cameras and microphone/speaker interactions can be captured to provide a time-aligned corpus.

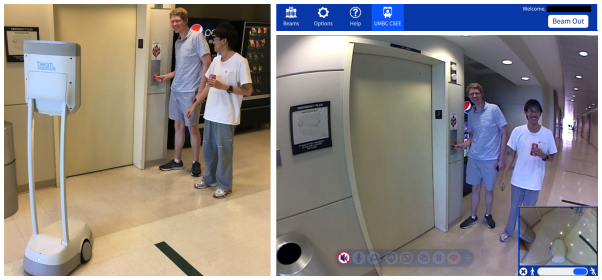

Figure 1: Telepresence-based context and language. *(left)* Bystanders push a button in response to a verbal request from the robot. *(right)* Video feed from the robot's sensors. In this example, the most visually salient elements of context are the elevator and people.

(3) End Users When a sufficiently robust model exists, we will instrument wheelchairs of proposed end users (e.g., ALS patients). This sensor array must be unobtrusive and relatively low power. This will include one or more time-of-flight Kinect 2 RGB-D cameras, an omni-directional powered microphone, and a high-resolution camera.

2.2 Context Interpretation

While any feature of as environment or actions may provide important context, we will focus on gathering sensor observations describing the most salient elements of the environment. We expect this to be primarily: 1) People in the environment, who will circumscribe the set of things the user is likely to want to say; 2) Objects in the environment, including fixed objects such as elevators; and 3) The environment itself.

Identifying elements of a scene is a difficult problem; initial efforts will use crowdsourcing or other low-cost, high-speed annotation of sensor data, but the broader intention is to use recent work on automatic identification of important visual elements (Carl Vondrick, 2016) and semantic labeling. (Anand et al., 2012)

Existing efforts on building language models from observations collect and train on corpora that are targeted to a particular scenario. Because we are gathering ongoing speech in a variety of settings, we are trying to learn from non-targeted speech, where the connection between the language and the sensor observations may be tenuous or non-existent. (For example,

a person may be talking about medication side effects while navigating.) Gathering data over a long period should allow irrelevant context to be weighted down in the learned model.

2.3 Language Learning

Our approach to text prediction inverts an existing model (Matuszek et al., 2013), in which the authors trained a joint model of language and context, and then treated language as a query against a perceived world state. In that work, the goal is to find a set of groundings in the world referred to by language x. The induced model is then $P(G|x, C)$, given data of the form $D = \{(x_i, C_i, G_i)|i = 1...n\}$, where each example i contains a sentence x_i, the world context C_i, and the actual referent (grounding) G_i.

In this work, we treat perceptual context as 'input' and language as 'output.' Given a similar dataset, the model to be induced is then $P(x|G, C)$. Our intention is to learn a similar model, incorporating semantic parsing and perceptual information and giving a probability distribution over possible generated outputs, as done elsewhere (FitzGerald et al., 2013). However, our initial experiments were performed using an n-gram prediction model.

The generation goal is a list of predictions from which a user can select. Since generated predictions can range in complexity from words to full sentences, generation strategies based on certainty can be applied, where more complex constructs are generated in higher-certainty situations. In this setting, providing a top-n list of results is useful, reducing the need for finding the single best generation.

3 Preliminary Study

For the preliminary user study, a set of four scenarios were shown to a group of fifteen participants. The scenarios asked each participant what they would say in each of four situations: a social interaction; answering questions from a doctor; asking someone to push an elevator button; and asking someone to retrieve a water bottle. The context was described to participants in text, simplifying out the question of how to represent real-world context. (See box for an example scenario and some responses.)

You have been having stomach pains after eating each day for the past week. You are visiting your doctor, who asks how you are doing. What is your response?

- "My stomach has been bothering me after I eat."
- "My stomach hurts whenever I eat."
- "I'm ok but I've been having stomach issues."
- "Good aside from the gut pain I'm having after eating."
- "I have been having stomach pains after eating each day for the past week."

3.1 Prediction Experiments

An interface was developed to test four different methods for generating predictions, of which three are novel to this work. These methods vary in the length of generated predictions: users were presented with combinations of single words, short phrases, or full sentences (see Table 1). A simulated QWERTY keyboard was available for fallback text entry. A new pool of participants were asked to use the interface to communicate responses to the same four scenarios, rather than typing responses on a keyboard.

In order to generate a predictive language model that is context driven based on these scenarios, n-gram models were constructed using the Presage predictive text entry program[1]. Four different prediction methods were tested using this model (see Table 1).

Method	Corpus	W.	P.	S.
StdEng	Standard English	✓	✓	
ContWord	Contextual	✓		
Context	Contextual	✓	✓	
ContSent	Contextual	✓	✓	✓

Table 1: The four text prediction methods tested, which vary in whether they generate words (W), Phrases (P), and sentences (S), and whether they are based on an existing English corpus or a preliminary contextual corpus.

For each participant/scenario pair, the number of selections (clicks) necessary to communicate was recorded. After each participant completed the tasks, they filled out a survey about the usability of the interface, how effectively

[1] `http://presage.sourceforge.net`, 2016-08-01

they felt they were able to communicate, and the perceived speed of each entry method.

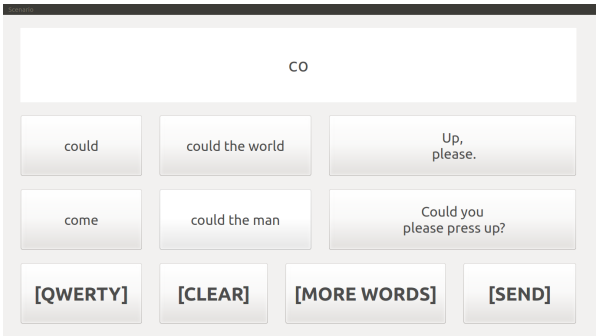

Figure 2: The prototype interface used in the preliminary study. The interface provides various options for text selection including full sentences and a virtual keyboard.

This pilot supported the hypothesis that users communicate faster with context-sensitive prediction (Figure 3); of the most-comparable methods, CONTEXT was faster than STDENG. While communication is fastest when complete sentences are shown, the users did not qualitatively prefer this option, underscoring the importance of personalized communication.

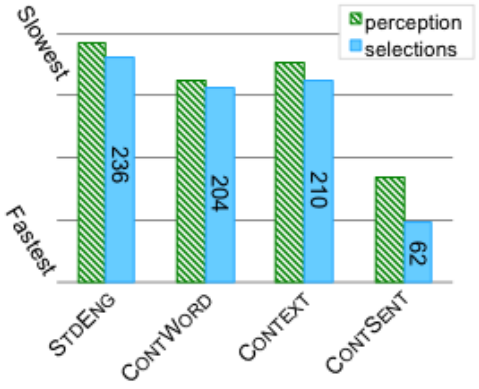

Figure 3: Participants' *qualitative* perception of the relative speed of different methods (green), compared to the number of selections actually used (blue). Perceived speed is shown as a weighted average of non-numeric rankings, and aligns closely with the number of selections required to complete a task.

4 Discussion and Future Work

We intend to pursue further experiments using more complete language and grounding models. For this, some simplifications must be addressed. The most immediate are the best way of modeling language and incorporating real-world context; this is necessary to know whether building a semantically informed, context-aware prediction model will present large gains in accuracy and acceptability. We believe this work will be able to contribute to the research community, providing leads and methods for more intelligent and usable language models. Nonetheless, while ambitious, our initial results support the belief that this approach has promise for text prediction and context-aware generation.

References

Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. 2012. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, page 0278364912461538.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics (TACL)*, 1:49–62.

Antonio Torralba Carl Vondrick, Hamed Pirsiavash. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

David L Chen and Raymond J Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37:397–435.

David L Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI*, volume 2, pages 1–2.

Juan Fasola and Maja J Mataric. 2014. Interpreting instruction sequences in spatial language discourse with pragmatics towards natural human-robot interaction. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 2720–2727. IEEE.

Nicholas FitzGerald, Yoav Artzi, and Luke S Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *EMNLP*, pages 1914–1925.

Zhiyu Huo and Marjorie Skubic. 2016. Natural spatial description generation for human-robot interaction in indoor environments. In *2016 IEEE*

International Conference on Smart Computing (SMARTCOMP), pages 1–3. IEEE.

Hadas Kress-Gazit and Georgios E Fainekos. 2008. Translating structured English to robot controllers. *Advanced Robotics*, 22:1343–1359.

James MacGlashan, Monica Babes-Vroman, Marie desJardins, Michael Littman, Smaranda Muresan, Shawn Squire, Stefanie Tellex, Dilip Arumugam, and Lei Yang. 2015. Grounding english commands to reward functions. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2013. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June.

Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2014. Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In *Proceedings of Robotics: Science and Systems*. Citeseer.

Raymond J Mooney. 2008. Learning to connect language and perception. In *AAAI*, pages 1598–1601.

Stephanie Rosenthal, Sai P Selvaraj, and Manuela Veloso. 2016. Verbalization: Narration of autonomous mobile robot experience. In *26th International Joint Conference on Artificial Intelligence (IJCAI16). New York City, NY*.

Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems, Berkeley, USA*.