Representing dynamic context dependencies in spatial prepositions as used in natural language indoor scene descriptions and object localizations

Stacy Doore

Abstract

Improving natural language descriptions of scenes and spatial localization of objects in indoor environments is a critical to creating effective dialogue-based indoor navigation systems (Kruijff, 2007; Cuayahuitl et al., 2012). Recently, Elahi et al. (2012) have suggested that mapping context based natural language description classifications directly to concepts in a linguistically motivated spatial ontology is sufficient to generate effective natural language navigation expressions. However, this approach of mapping these classifications directly to these concepts, may be oversimplifying important aspects of spatial prepositions and their context dependencies based on assumptions made about the relative weight given to the spatial configuration vs. functional interaction characteristics and the nature of the relationships between these two factors. This paper explores how the weight of multiple factors is manifested in two different corpora in order to determine what role specific aspects of spatial prepositions semantics play in different types of indoor settings or tasks. The scope of this study is spatially situated within indoor settings at the scale of a room description level (vs. the table top level) and is focused solely on indoor scene descriptions and object localizations used for navigation tasks.

'Setting the scene'

Right now, millions of people are using intelligent dialogue-based navigation systems to find their way around and complete ordinary tasks.

"Diri, how do I get home from here?"

"Well Stacy, take left onto College Avenue, then at the left turn left onto Stillwater Avenue…."

Diri, I want a cup of coffee. Is there a Starbucks on route to home?"

"Yes, there is a Starbucks located on Bangor Mall Blvd. You will need to get off 1-95 at exit 187 and then bear right onto Hogan Rd."

But, using the same kind of mobile device system to navigate around in indoor space or trying to use it to complete tasks has not quite reached the same level of development.

For example, I might be in an unfamiliar grocery store, and want to ask,

"Diri, where is the orange juice?"

In this simple request, I am asking the system to locate my location, look up a map from the store's public API, and then provide step by step directions to the juice, updating my position as I move.

Or a friend of mine who is blind might be at a conference in a huge hotel, and needs to find her way to the reception desk but can't use landmarks, signs or other visual clues to navigate. She might like to ask her personal assistant,

"Diri, I need to check into my room, how do I get to the reception desk? Here is where I am right now."

She takes a picture of her current location view to identify her position with her phone.

Seems a simple enough request, right? Wrong.

Unlike an outdoor space navigation system that works off of a big map and a search box, an indoor navigation system must be able to link with dynamic data and receive cues. In some cases, this might come from the buildings themselves which are equipped with the ability to locate any Wi-Fi signal to estimate the position of the device. However, the system may not be as accurate as we would like and can improve its accuracy by having the user give verbal or in this case visual landmarks within the environment.

Such a system would need to access multiple ontologies to locate the user in physical space, interpret the image for visual clues as to the user's position, interpret the user's natural language query for meaningful cues as to the intended task the user wants to accomplish, access any public APIs to retrieve the information required to complete the task and then provide step by step directions to reach the desired destination. In order for this to work, the system needs to be fed by extremely refined ontologies able to interpret the semantics of the user's navigation and task requests, never mind the image processing and interpretation of the image into natural language to provide a description of the users current location.

The scope of the research is confined to only a small slice of this problem: translating a conceptual spatial model into a natural language expression, using ontologies for the translation task. While a spatial linguistic ontology exists which is grounded in theory of indoor space, we believe it may not account for dynamic spatial context dependencies. More specifically, the representation of specific aspects of context dependencies in spatial prepositions- physical and perceptual access, order of potential encounter, and direction based on general or lateral orientation- is critical to accurately translating a conceptual spatial model into natural language expressions that fits dynamic user goals and tasks.

Spatial prepositions: use and characterization

This study is an investigation into Langacker's ideas about resolving the "basic question" posed by Herskovits (1980) and Vandeloise (1985; 2006) regarding the nature of spatial prepositions, which in some cases can be characterized strictly as an expression of 'spatial configuration' while in other cases, they might more accurately be described as 'functional interaction' (Langacker, 2010). Langacker's approach also addresses the "alignment question", which asks, if these multiple characterizations of a preposition can be determined, can spatial and functional

factor relationships be sufficiently represented to reflect true nature of a preposition's implicit semantics?

A body of cognitive linguistics research has found that contextual dependencies provides weighting as to how spatial prepositions are used and understood in stable settings, their use and meaning in 3D scene descriptions and dynamic wayfinding tasks is less clear. Although, Langacker concludes that in most cases 'spatial configuration' can be characterized as the primary characterization of prepositions over the 'interaction function', he acknowledges that less weight is given to capturing the semantics of prepositions that "pertain to human interaction with the world at the physical, perceptual and purposive levels". This is referred to as "anthropomorphic" characteristics (Vandeloise, 1986: 119) or "embodiment" in cognitive linguistics (Lakoff, 1987). The *complex* nature of many spatial prepositions (i.e. the exact weight ascribed to its 'spatial configuration', 'interactive function' properties) is present when we observe:

The box is *in front of* the chair.

The lamp is *by the side* of the bookcase.

The room is *at the top* of the stairs.

The prepositional phrase *in front of* denotes a general orientation which is defined by coincidence of line of sight, the direction of anticipated motion, and the frontal orientation of the subject. Langacker concludes that there are three major functional entities that comprise the conceptual characterization of a spatial preposition: the **trajectory** which functions as the *target* or the entity one might be trying to locate, the **landmark** which functions as the *reference point* or the entity one uses to find another object. Finally, there is the **search domain** or the limited region within which the target can be found (Langacker, 2010).

The examples given above illustrate the complex conceptual meanings of prepositions which combine spatial configuration and interactive properties, whose importance might vary depending on the preposition and how it is used. Therefore, in an indoor space at the scale of the room level it might be important to include the anthropomorphic aspects such as the physical and perceptual access, the order of potential encounter, and the direction based on general orientation in the classification of a complex preposition. Would the weighting and its consistency of use of these aspects convey a more accurate and precise interpretation of meaning and provide more effective directions during indoor navigation tasks? Is using the most linguistically appropriate or correct spatial preposition important in some tasks but not others? Does the use of more general spatial prepositions misrepresent the search domain in indoor space in some cases but not others? What are the contexts where it is critically important to use more precision in an indoor setting?

GUM-Space ontology

While there are many foundational ontologies, which are used to convey semantic descriptions of spatial concepts (e.g., BFO and DOLCE), the spatial extension of the Generalized Upper Model (Bateman et al., 1995; GUM-Space, Bateman et al., 2007) formalizes categories that are relevant for natural language of space (Bateman et al., 2010; Hois and Kutz, 2008a; 2008b). It builds on research in cognitive linguistics (Talmy, 2006; Levinson, 2003; Halliday and Matthiessen, 1999),

and on analysis of natural language corpora. The primary aim of GUM-Space is to provide a coherent representation of spatial language as an underlying linguistic basis for dialogue systems. The GUM-Space extension refines linguistic components necessary to formally specify spatial language utterances (Bateman et al., 2007). It is intended for use within natural language dialogue for spatial assistance (Ross et al., 2005) and for interpreting spatial expressions in a situational context, for instance, in relation to formal representations of *spatial scenes* (Tyler and Evans, 2003; Hois and Kutz, 2008a; 2008b).

GUM-Space concepts and relations allow for the representation and analysis of spatial utterances for their spatial meaning by spatial location, position, positioning, orientation, or movement of entities, spatial relations between entities, spatial change, and groups, combinations, or connections of the given spatial information (Hois, 2010). Gum-Space provides approximately 70 difference types of spatial relations (SpatialModality), which define how entities can be located in space. For example, the sentence, "The chair is next to the table.", GUM-Space would categorized the elements as:

SpatialLocating : *(ontological class)*
locatum: chair *(instance of the class Locatum)*
processInConfiguration: is
placement: GeneralizedLocation : *(ontological class)*
hasSpatialModality: RelativeNonProjectionAxial *(ontological relation)*
relatum: table *(instance of the class Relatum)*

Gum-Space is also designed to provide formalized information on entity motions, orientations, routes, and directions by breaking down the static and dynamic spatial units in a sentence into different types of configurations (Hois, 2010).

Current study rationale and methods

GUM-Space has been evaluated for its inter-annotator reliability and its spatial logics using a number of spatial language corpora (Hois, 2010; Hois and Kutz, 2011, Elahi et al., 2012) such as the Trains 93 Dialogues (Heeman and Allen,1995) and IBL (Instruction Based Learning) (Lauria et al., 2001), the HCRC Map Task (Anderson et al., 1991) and most recently, the CReST corpus (Eberhard et al., 2010). These studies have found that while GUM-Space is adequate for structuring spatial language so that non-experts are able to understand and use the complex annotation schema of GUM-Space, there was some confusion when evaluators were faced with similar but slightly different annotations, specifically, those categories that are specified hierarchically close together, but need to be considered in context. These studies also asked their annotators to 'clean up' the 100 sample sentences in the corpora, and remove the non-spatial data because it could lead to different interpretations.

We believe that the results of these previous studies provide sufficient motivation and justification to analyze a number of spatial linguistic corpora (e.g. the CReST corpus and the VEMI corpus (Kesavan and Giudice, 2012) to look for evidence of relative weights given to spatial and functional factors of spatial prepositions and evaluate at how well GUM-Space might represent the more *anthropomorphic* aspects of spatial prepositions in a strictly indoor setting given a number of different task scenarios. While previous studies mapped CReST corpus

sentences to GUM-Space (Eberhard, 2010; Elahi et al., 2012), each was more focused on validating GUM-Space concepts to show that it could adequately represent spatial semantics to localization expressions regardless of contextual interpretations. Neither study considered how well GUM-Space might weight the importance of spatial as primary over function factors based on the context of the utterance and/or response. Nor did they consider how and when the representation of specific aspects of context dependencies in spatial prepositions - physical and perceptual access, order of potential encounter, and direction based on general or lateral orientation- might be critical to accurately translating a conceptual spatial model into natural language expressions that fit different user goals and tasks.

We also believe that the CReST corpus should lend itself quite well to an analysis of these factors (for a full description of the CReST corpus, see Eberhard et al., 2010). In addition, we will use a more recent spatial linguistic corpus, the VEMI corpus. This corpus of direct observations vs. photographic observations of an indoor scene in a virtual reality environment, consists of 16 participants per room and 16 descriptions per image (n=32 descriptions). The preliminary analysis of this corpus concluded there were no significant differences between the subjects' descriptions of the scenes in the photo vs the VR environment, the subjects were found to use different strategies in describing the scene (e.g. center to side strategy vs. going around room) and there were some advantages of certain types of strategies over others (Kesavan and Giudice, 2012). Both corpora will be analyzed using similar methods to previous studies (Eberhard, 2010; Elahi, 2012) such as: 1) POS tagging using C7 tag set (for richer descriptions); 2) Use NLP (desktop and online) tools for generating root trees and dependency annotations; 3) Code corpus for Dialogue Structure Annotation, Disfluency Coding, POS Annotation, Constituent Annotation, Dependency Annotation, and 4) map results to GUM-Space concepts and relations.

Conclusions

The goal of this study is to learn if an existing spatial ontology can be used to generate scene descriptions and natural language expressions for dialogue-based indoor navigation systems that can represent of specific aspects of context dependencies in spatial prepositions. We also will investigate how and when these more precise representations might be critical to accurately translating a conceptual spatial model into natural language expressions based on different user goals, tasks, and needs. In determining the more precise use and meaning of spatial prepositions in different contexts, we hope to offer insights into how design more effective and accurate multimodal indoor navigation systems that offer users a wider variety of tools with which to understand and interact with the world around them.