Exploiting Semantic and Pragmatic Information for the Automatic Resolution of Spatial Linguistic Expressions

Andrea Corradini
Computational Linguistics Department
University of Potsdam, D-14476 Golm, Germany
andrea@ling.uni-potsdam.de

Abstract

We present a computational model for the interpretation of linguistic spatial propositions in the restricted realm of a puzzle game. Based on an experiment aimed at analyzing human judgment of spatial expressions, we establish a set of criteria that explain human preference for certain interpretations over others. Each criterion is associated to a metric that combines the semantic and pragmatic contextual information regarding the game as well as the utterance being resolved. By resorting to machine learning techniques we determine a model of spatial relationships from the data collected during the experiment. Sentence interpretation occurs by matching the potential field of each of its possible interpretations to the model at hand. The system's explanation capabilities lead to the correct assessment of ambiguous situated utterances for a large percentage of expressions.

1 Introduction

The interpretation of spatial expressions is an important aspect of human cognition. Several experimental and theoretical studies have analyzed how language is linked to the non-linguistic spatial world with the goal to shed some light on the human mental processes that underlie the understanding of linguistic utterances involving space. Findings from these research endeavors have paved the way for the development of computational systems able to analyze, interpret and generate natural language descriptions of space and the physical world.

In this work, we focus on the interpretation of three types of linguistic relationships that form the basis for spatial expressions: topological relations like "*near*", projective relations such as "*left of*", and the relation "*between*". Projective relations need the specification of a frame of reference.

Within the scenario of a speech-operated 2D puzzle game, we have been developing a computing system able to understand the meaning of and consequently act upon linguistic instructions like e.g. "*land the green piece over the T-shaped one*" that can be ambiguous to a human who is not embedded in the same situation and sharing the same conversational context of the speaker/writer.

The paper is structured as follows. First, we discuss relevant related works. We then present the motivation for this research and the computational model that we developed based on the experiments we carry out. Eventually, we propose a system evaluation and a discussion on future extensions.

2 Related Work

Researchers in the field of language-oriented artificial intelligence have proposed several methods to deal with the inherent ambiguity of language and to handle traditional linguistic phenomena like presupposition, quantification, anaphora, under specification, and elliptic expressions. In parallel to research on these well-known sources of ambiguity, the understanding of propositions that depend on situational context has emerged as an active area of study and the treatment of spatial information in utterances has evolved into an ever growing field.

A relevant number of conceptual models that relate language to visual spatial information have been proposed (Eschenbach 1999; Tapus et al., 2005). Backed by theoretical works and/or empirical experiments (Costello & Kelleher, 2006; Logan & Sadler, 1996), more and more computational models that exploit the potential of verbal communication to interact with visual or spatial data have been implemented particularly for natural language interfaces to graphical systems and human-robot interaction.

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 67–70,
Antwerp, September 2007.

The SHRDLU system (Winograd, 1971) is probably the first relevant work that shows how syntax, semantics, and reasoning about the world can be successfully combined to produce a system that understands natural language to control the actions of a simulated robot arm. Following this pioneering work, other prototypes and models have been put forward for topological and projective relations. Several works based on language modeling and visual context (Gorniak & Roy, 2004; Roy et al., 2002; Roy & Mukherjee, 2005) involve aspects of grounded situation model. These approaches lead to the development of visual context sensitive grounded systems that understand, learn and generate natural language. A research methodology that addresses common problems in spatial communication arising during human-robot conversation is outlined in (Moratz et al., 2001). In (Kelleher et al., 2005) visual information, context and salience are integrated to leverage the understanding and generation of spatial expressions in the context of virtual reality applications. A variety of metrics and potential field measures are introduced in (Kelleher et al., 2006; Regier & Carlson, 2001) as a powerful tool to model and characterize spatial relations among 2D objects as perceived by human subjects. An integration of potential field models with visual information to control a robot that follows natural language commands to perform manipulative actions is presented in (Brenner et al., 2007) for the task of action planning in situated communication. In (Gorniak & Roy, 2005; Gorniak et al., 2006) the use of situated communication in computer games is investigated.

Excluding (Roy et al., 2002), the works outlined above have not resorted to machine learning techniques. Our work shares with (Kelleher et al., 2005; Kelleher et al., 2006; Regier & Carlson, 2001) the idea of encoding spatial information using a set of local metrics. It differentiates from them in the way we perform the assessment of the values of the metrics.

3 Resolving Spatial Expressions

3.1 Situated Communication in Pentomino

Pentomino is a popular recreational math puzzle game. The game consists of twelve different pieces that are built as arrangement of five square units joint along their edges. The objective is to fill up a given game board using all pieces. To accomplish this task, players can select, rotate, translate, flip, remove, mirror, and land pieces onto the board. In early studies on human-human communication to play Pentomino, we noticed that subjects resort extensively to localization expressions when they intend to collaboratively resolve a puzzle thus making this game an excellent prototyping arena for situated natural language understanding.

Our model is integrated into a digital version of Pentomino where speech can be used as a complementary input mode (Corradini et al., 2007). We exploit the game semantics and pragmatic along with context information available from both the visual display on the user interface and the game history to interpret spatial expressions used to play the game. At anytime, the player is allowed to customize a few application settings that affect the visual feedback and thus in turn visual-grounding (Roy et al., 2002; Roy & Mukherjee, 2005) of context information that bridges the symbolic realm of linguistic concepts with entities in the game world.

3.2 An Experimental Study

To investigate human interpretation of spatial situations, we run a psycholinguistic experiment that parallels the task of an automated system for playing Pentomino. We collected data from 38 participants (22 males and 16 females) both native and non-native English speakers with age ranging from 13 to 72 years ($\mu = 31.3$, $\sigma = 13.5$). Subjects were given a set of 40 image-text pairs and instructed about the game objective and rules. We showed the subjects a snapshot of a puzzle game and the next instruction to carry out in text format as a single separate instruction. Subjects were then asked to update the board according to their interpretation of the instructions with the goal to maximize the possibility to finish the game after carrying out the move. We chose such a setting both to elicit controlled spatial interpretations in different situations and to collect data that can give insights on factors, motivations, and mechanisms that play a role in turning the mental picture of a linguistic sentence into an actual spatial configuration.

A post-study analysis of the corpus of 1520 task solutions showed that while all subjects implicitly used themselves as frame of reference (see Figure 1) a few different configurations were proposed for each single task. One annotator searched for pragmatic and semantic errors in the solved tasks. We considered as a pragmatic error any spatial ma-

nipulations that, once performed, would at once appear to lead to no game solution i.e. result in the creation of one or more islands of cells with less units than the number of squares making up a single Pentomino piece. We refer to these small holes onboard to as *smHoles* (see Figure 1). We classified as semantic errors all cases of spatial actions and instructions that violate the game rules or were impossible to carry out. A second annotator scored 24 randomly selected user forms i.e., a 63.2% random sample. Compared to the first annotator there was a 98% match on what the error events were. In total, we found an average 8.3% of pragmatic errors ($\mu = 4.8$, $\sigma = 4.6$) and a negligible 0.02% ($\mu = 0.6$, $\sigma = 1.5$) of semantic errors. After removing these error cases from our corpus, we analyzed the remaining 1394 picture-instruction couples (91.7% of the data) to infer a best estimate of the space considered by the subjects given a spatial relation among reference objects.

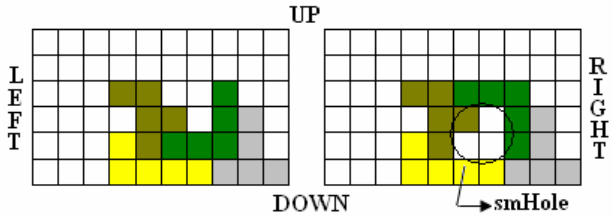

Figure 1. (left) A correct semantically and pragmatically interpretation of the instruction used in Section 1; (right) a pragmatically incorrect one. Text around the borders indicates the implicit frame of reference.

The computational model we developed bases on both the analysis of the data collected and the fact that in the context of a restricted language and limited number of visual entities, subjects tend to refer to objects by listing their properties and attributes such as color, shape and size (Roy, 2002).

3.3 Criteria & Metrics

From data of our experiment, we realized that for relations of the kind "*near*", "*under*", "*left to*" etc., over 97% of the subjects considered locations on the board grid that are within a certain small distance to the referent. In the case of "*between*" relations, 87% of the subjects considered points at locations mid-way to the referents. According to the relation at hand, we refer to the area including the points that satisfy the proximity requirement as *region of interest* or *RoI* in short. It restricts the set of possible locations referred to in the utterance.

We define a series of metrics over the *RoI* based on the notion of field potential (Kelleher et al., 2006). They describe degrees of likelihood of acting upon an object at a given location according to a set of criteria that capture and incorporate the most commonly used interpretation strategies adopted by subjects of our experiment. Given a sentence that refers to object *Obj* via a spatial relation *Rel* to another reference object *Ref*, they are motivated by the observation that people tend to:

C1) operate on *Obj* that is as closer as possible to *Ref* (*Proximity criterion*)

C2) operate on *Obj* at positions that maximize the number of physical contacts with other game entities such board edges or other pieces (*Adherence criterion*)

C3) operate on *Obj* at positions that maximize the intersection area between *Obj* and the *RoI* (*Communality criterion*)

C4) operate on *Obj* at positions that either minimize distance between *Obj*'s and *Ref*'s centers of mass or, in case of a "*between*" relation, are equidistant from those of all other referents (*Center of Mass criterion*)

C5) Play uniformly i.e. they concentrate on a region on the board which try they fill in incrementally before moving to other distant areas of the board (*Location Saliency criterion*)

C6) Avoid the creation of *smHoles* since they make the game unsolvable (*Fillability criterion*)

The criterion C6 captures aspects relative to game pragmatics and semantic knowledge. Criteria C1 to C4 reflect game's geometrical considerations at a given time. The criterion C5 accounts for the dialogue context in terms of game history. For each criterion we defined a corresponding metric to quantify its salience value at a specific location.

3.4 Spatial Expression Resolution & Results

Anytime a spatial utterance is processed, we try to carry out the underlying instruction at each point in the *RoI*. If this is possible, we then calculate the normalized metric values on those points. We thus have a kind of field potential whose intensity is modulated by the degrees of likelihood of each criterion after the particular instruction is executed at a given location. To select the correct placement,

we use multiple linear regression to model the relationship between these likelihoods and an expected response variable depending on the location by fitting a linear equations to the observed data. The model is defined by the k parameters $\beta_1..\beta_k$ of the system of linear equations:

$$Y_i(P) = \beta_0 + \beta_1 f_{i,1}(P) + \ldots + \beta_k f_{i,k}(P) \qquad (1)$$

Here k is the number of criteria, $f_{i,k}(P)$ the values (the independent variables) of the metrics applied at location P in the *RoI*, $Y_i(P)$ the expected goodness value (the dependent variable) at P, i an index running over the number of possible placements of the piece being manipulated and for each of the 5 units making up that piece. In our model, $Y_i(P)$ is set to 1 for all units P of the piece 1 if its manipulation can be found in our corpus of human interpretations, to 0 otherwise. Ultimately, the values β_j act as weighing coefficients for the metrics' values. We use equation (1) as a combined likelihood to gauge how close a spatial configuration is to the model of human interpretations. Specifically, we rank any location in the *RoI* according to the value obtained by summing up equation (1) over each point of the piece after this is operated upon.

We used half of the data for the determination of the model parameters and half for the evaluation. By taking the maximum value of the ranked list, the model interpreted spatial descriptions as humans did in our experiment in 61.4% of the expressions. Correct interpretations were ranked either second or third in 16.2% of the cases.

4 Discussion and Conclusion

We implemented a computational model that attempts to approximate human interpretation and judgment of situated language in the micro-world of a 2D puzzle game. We believe that the probabilistic nature of our method can be very useful in a dialogue system for spawning clarification requests or suggesting the location for a certain instruction.

Our system confirms that adopting an approach that considers several sources of information such as context, semantic and pragmatic evidence can be beneficial to the understanding of situated utterances (Gorniak & Roy, 2005). The metrics, now tailored for our restricted game domain, are extendible to other grid-like scenarios and spatially aware systems, even in 3D. The resolution of spatial relations is also portable to the case of one-to-many relations by applying our strategy between the one object and each one of those in the group.

We are expanding the system to include a few more metrics and dialogue capabilities between player and system, for error resolution and in contexts that need clarification to resolve ambiguities.

Acknowledgement. The EU Marie Curie grant #FP6-2002-Mobility-3014491 supported this work.

References

Brenner, M., Hawes, N., Kelleher, J., and Wyatt, J. 2007. *Mediating between Qualitative and Quantitative Representations for Task-Oriented Human-Robot Interaction.* Proceedings of the IJCAI.

Corradini, A., et al. 2007. *A Robust Spoken Language Architecture to Control a 2D Game.* Proc. of AAAI Int'l FLAIRS, pp. 199-204.

Costello, F., and Kelleher, J. 2006. *Spatial prepositions in context: The semantics of near in the presence of distractor objects.* Proceedings of the ACL-Sigsem Workshop on Prepositions.

Eschenbach, C. 1999. *Metric Details for Natural Language Spatial Relations.* ACM Transactions on Info. Systems, 16:(4):295–321.

Gorniak, P., and Roy, D. 2004. *Grounded Semantic Composition for Visual Scenes.* Journal of AI Research, Vol. 21, pp. 429-470.

Gorniak, P., and Roy, D. 2005. *Speaking with your Sidekick: Understanding Situated Speech in Computer Role Playing Games.* Proceedings of AI and Digital Entertainment.

Gorniak, P., Orkin, J., and Roy, D. 2006. *Speech, Space and Purpose: Situated Language Understanding in Computer Games.* Annual Meeting of Cogn. Science Society Workshop on Computer Games.

Kelleher, J., Costello, F., and van Genabith, J. 2005. *Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context.* Artificial Intelligence, Special volume on connecting language to the world, 167(1-2):62-102.

Kelleher, J., Kruijff, G.J., and Costello, F. 2006. *Proximity in Context: an empirically grounded computational model of proximity for processing topological spatial expressions.* Proc. of COLING.

Logan, D. and Sadler, D. 1996. *A Computational Analysis of the Apprehension of Spatial Relations.* Language and Space, MIT Press.

Moratz, R., Fischer, K., and Tenbrink, T. 2001. *Cognitive Modeling of Spatial Reference for Human-Robot Interaction.* International Journal on Artificial Intelligence Tools, 10(4): 589-611.

Regier, T., and Carlson, L. 2001. *Grounding spatial language in perception: An empirical and computational investigation.* Journal of Experimental Psychology: General, 130(2):273-298.

Roy, D. et al. 2002. *A Trainable Spoken Language Understanding System for Visual Object Selection.* Proceedings of the ICSLP.

Roy, D. 2002. *Learning Visually Grounded Words and Syntax for a Scene Description Task.* Computer Speech and Language.

Roy, D., and Mukherjee, N. 2005. *Towards situated speech understanding: Visual Context Priming of Language Models.* Computer Speech and Language, 19(2):227–248.

Tapus, A., et al. 2005. *Towards a multilevel cognitive probabilistic representation of space.* Proc. of the SPIE, Vol. 5666, pp. 39-48.

Winograd, T. 1971. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.* MIT AI Technical Report 235.