Tailoring the Interpretation of Spatial Utterances for Playing a Board Game

Andrea Corradini

Institute of Business Communication and Information Science
University of Southern Denmark
6000 Kolding, Denmark
andrea@sitkom.sdu.dk

Abstract. In order to build an intelligent system that allows human beings to cooperate with a computing machine to perform a given task it is important to account for the individual characteristics of each user. In this work, we describe a flexible and adaptive user interface that is capable of interpreting spatial linguistic expressions according to the peculiar preferences and use of language of the user. As a proof of concept, we have implemented a graphics board game in which players can move, remove and spatially manipulate in 2D various game pieces on the computer screen by issuing speech commands during a dialogue session with the application. After an initial guided interaction, the system analyzes on the fly subjective interpretations of several linguistic relationships that form the basis for spatial expressions to tune a set of system’s parameters. We tested the system with different human subjects. Automatic judgment of spatial expressions based on each user’s interaction behavior over-performed sentence interpretation of a model that we previously created to tailor the characteristics of some abstract typical player. Post interaction informal talks also indicated a higher level of user satisfaction in subjects playing with the customized application rather than the more general one.

1 Introduction

One of the results of the growth in information and communication technology research is the large quantity of electronic devices available nowadays at a reduced cost. Low-cost digital equipment has boosted the number of people adopting new technologies, thereby causing an increased diversification of end users. The inherent user variety poses several issues for designers and developers who want to create robust, reliable and usable systems that accommodate to the individual needs and distinct characteristics of a large number of people. The demand for personalized applications also gives rise to a need to expand the range of tasks and the physical contexts in which perform them.

The standard dyadic model of human-computer interaction for knowledge-based systems assumes that machines process user input according to a kernel of knowledge [10]. This body of knowledge ranges from information about the application domain to constrain user actions and available system operations [14], to information regarding the communication process to determine whether, how, and when to assist the
user [15], and finally to knowledge about the current communication partner for the system to interact with the user in a cooperative way. Any successful man-machine interaction strategy depends on both the context at hand and the current user for there is no typical system user. Existing research has thus attempted to address the challenges posed by adaptive application by focusing on how content can be customized on the basis of user models and interaction history.

In this paper, we present a multimodal interface in which speech, typing, and mouse clicking can be used to control a run-time board game system. We exploit the information from the different input modes to interpret user’s intention and solve ambiguities in spatial utterances containing topological relations like “next to”, projective relations like “under”, spatial relations of the kind “between”, and absolute indication of spatial references such as “upper rightmost corner”. We exploited the behavioral data collected from players interacting with the game to infer a set of system parameters to model the average player. In an optional introductive session, new players are asked to follow a protocol to learn how to interact with the game. During such a session, the system accommodates for each user’s playing behaviors by adjusting its internal parameters while building a new one that is personalized to the subject currently playing.

In the rest of this paper, we start with a review of works that relate to our approach. We then outline the main features of our systems. We continue by presenting the results of a usability study accomplished with human subjects to check the feasibility of our method. Finally, we conclude with a brief discussion and notes on issues of user modeling in our context.

2 Related Work

The end users’ diversification is tautological to say that there is no universal interface and thus calls for user applications that can accommodate to the given user, task and context at hand. Over the last years, several interfaces have been proposed to take advantage of the individual user differences in education, environment, physical impairments, cognitive abilities, social background, age, etc. The search of a strategy that helps determining recurrent patterns in user behaviors and further exploits them to generate an appropriate contextualized response is a central task [24] common to most of these systems.

In early works, adaptable interfaces were put forward to achieve a certain degree of system personalization. This was accomplished typically by requiring the user to either set some preferences in advance or fill in some form from which a designer could infer a set of preferences [1]. State-of-the-art adaptive systems require approaches that are more complex. The aim of those systems is to ferret out what is in the head of the user automatically from the analysis of off-line interactive sessions i.e. with data collected in advance and/or on the fly as the user interaction proceeds [9]. A detailed review of the development of user modeling systems is given in [18]. [31] presents a study of issues in user modeling with focus on standard machine learning techniques.

As pointed out in [3], sometimes the terms context-awareness and adaptation are used interchangeably. However, they denote distinct concepts where the former one
refers to the digital device capability of accommodating to the physical-virtual context in which the human user is situated. This tightly relates to the notion of location and environment and is relevant for interactions with electronic equipments that operate in immersive environments and/or allow the user to be mobile [7, 13, 25]. Both user and device models must be accounted for to achieve customized content, presentation, and response.

Voice user interfaces offer a great potential for enhancing the interaction with computing machines. Due to the characteristics of human language, speech based applications are an ideal arena for experimenting with the concepts of adaptation and personalization. Over the last two decades, there have been a variety of practical task oriented spoken conversational systems for applications in limited domains [12, 16, 20, 27, 28]. These prototypes have provided good results in laboratory settings but have become commercial applications only in a few cases [22], notably in customer service centers and telephone-based information systems. Current technology limitations in addressing human language ambiguities and the difficulties to achieve user modeling in natural language applications [19, 32] are among the main reasons of this reduced performance in comparison with that of human beings. To this extent, some researchers have tried to improve the acceptance of dialogue systems by adapting the application grammar to the user’s use of language either by adding new words and assigning a semantic meaning to them [8, 17] or by allowing users to expand it at run time with new rules [11].

Systems that exploit complementary information from other modalities have been also proposed to enhance spoken communication through user adaptation [16, 21]. Such multimodal spoken and dialogue systems offer more alternatives for adapting to user preferences and needs and indeed have been shown to achieve a higher degree of satisfaction in their users [26]. More work has focused on how to customize content based on user models and interaction history [2, 6, 26] similarly to early research on adaptation in hypermedia.

All the systems mentioned so far are difficult to compare with each other for the lack of standardized evaluation procedures. Despite a methodology to test multimodal output generated in response to a specific set of user, device, and situation constraints has been presented in [23], a thorough evaluation of an adaptable/adaptive multimodal systems remains a difficult and challenging task.

3 The Adaptive Game System

3.1 The Setting

As a test-bed for our approach we chose the digital version of a math puzzle game called Pentomino. In such a game, each piece is composed of five congruent unit squares, connected orthogonally. Reflectional and rotational symmetries of Pentomino pieces do not count. Hence, there is a set of only twelve valid different arrangements that may be used to play. They are usually named after the letters of the Latin alphabet that they resemble. Solving a Pentomino puzzle consists of tiling a two dimensional game board (see e.g. Figure 1), i.e. covering and filling a designated region without overlaps and without gaps using up the available pieces. Before
placement, pieces may be rotated, shifted, flipped, and/or mirrored to make them fit onto the intended board location.

Early studies that we carried out on human-human communication to play the game show that subjects resort extensively to localization expressions when they collaboratively play towards the resolution of a puzzle. This makes the game an excellent scenario for testing ambiguities in the use understanding and interpretation of spatial language.

3.2 Resolution of Spatial Expressions for the ‘Average’ Player

In [4], we presented an approach for the creation of a computational model capable of interpreting a set of linguistic spatial propositions in the restricted domain of Pentomino. In an experiment with 38 subjects, we analyzed human judgment of spatial expressions that allowed us to come up with a set of criteria that explain human preferences for certain interpretations over others.

For each of these criteria, we defined a metric that combines the semantic and pragmatic contextual information regarding the game as well as the utterance being resolved. Each metric gives rise to a potential field that characterizes the degree of likelihood for carrying out the instruction at a specific hypothesized location.

We then resorted to multivariate linear regression techniques to determine a model that incorporates and summarizes all spatial interpretations given by the subjects during the experiment. By matching this model with the metric values calculated for an unknown spatial relationship, we could then resolve this latter to a specific location. The created model could be used to correctly explain spatial sentences of any hypothetical average user for 2.5 spatial sentences out of 4 on average.

This number indicates that there is still room for improvement.

3.3 Adaptation of the Resolution of Spatial Expressions to the Player

The analysis of data collected from observing human beings resolving spatial utterances while playing Pentomino, made it possible for us to come up with a set of nine criteria [4] that give helpful insights on the underlying strategy that people adopt to interpret situated communication in the game world. Each criterion reflects a specific behavioral pattern, which is common to the great majority of subjects. At the same time, each criterion incorporates some implicit parameters, which vary in quantity from individual to individual.

For instance, we noticed that users tend to place a piece Obj as closer as possible to the reference piece Ref to resolve sentences like “drop the green piece left to the blue one” (here being Obj the green piece and Ref the blue one). This proximity criterion generalizes very well over all subject population yet it tells us little about how we can quantify the notion of closeness in terms e.g. of number of units on the game board. For the average user, we can exploit the finding that over 97% of the subjects considered locations on the board grid that are within a distance of up to three units far away from the referent. This information restricts the search area for the resolution of the spatial relation under investigation since it indirectly defines an area of possible candidate locations on the board.
We refer to this area as the *region of interest, RoI* in short. Eventually, any ambiguous utterance involving spatial relations resolves to a location within the *RoI*.

If we further focus on each specific user, we can restrict such region and make it fit even more to the user’s own mental idea of closeness. Hence, the next step consists in defining a maximum and minimum value for the extension of the *RoI* that can quantitatively express the notion of closeness as intended by each specific user. In the quest for such a peculiar value, we have to factor in also other criteria that usually co-occur in the characterization of human interpretation of spatial utterances. For instance, we need to account for findings that indicates that users tend to place a piece *Obj* at positions that maximize the number of physical contacts with other game entities like e.g. board edges or other pieces onboard (*adherence criterion*). Moreover, we must take into account the evidence that preferred positions for the placement of *Obj* are those that minimize the distance between the centers of mass of piece *Obj* and the referent *Ref*, respectively (*center of mass requirement*). Other factors that have an influence on a personalized choice of the *RoI* can be inferred from the analysis of the game history. We observed uniformity in the way users play the game: they start building up the puzzle from a certain region and then incrementally make it bigger and bigger. They seldom jump from one area of the board to another one farther away from it (*uniformity criteria*). Players that want to pursue the game objective do not place pieces on the board in a random way except in very specific cases (Figure 1). The analysis of the game history helps restricting the possible locations to include in the *RoI* thus ultimately accommodating for the playing style of the user. Finally yet importantly, an experienced player knows that piece placement must not give rise to enclosed regions on board that cannot be filled up with the remaining pieces since this would make the puzzle unsolvable (*solvability criteria*). This criterion relies on contextual game conditions as well as game semantics and thus relates directly to the user experience and skills. This becomes more and more important as the player improves and becomes an expert player.

In order to determine a user-specific *RoI* underlying the criteria that we briefly touched upon previously, we guide new users through an interactive session where they are asked by the application to perform a sequence of specific game moves. For this task, we created approximately 500 different configurations to provide a broad comprehensive coverage of use cases for each criterion as well as for situations where more criteria are potentially in conflict with each other (Figure 1).

In the same way as outlined in [4], we then resorted to multivariate linear analysis to determine the parameters that act as weighting factors for each of the criterion. These parameters determined by this standard machine learning technique are tailored to the playing style, behavioral patterns, knowledge of the game, and skills of the current user ultimately constitute an average user model. At the same time, the analysis of the response of the subjects to the different tasks assigned to them, allows us to determine a set of regions of interests for each kind of spatial relationship personalized to the particular individual. The determination of a single *RoI* to explain any given spatial relation class occurs then by inter-class intersection of each of the *RoIs* as determined during the guided session. Spatial ambiguity is eventually resolved by matching each game move with the average user model as explained in detail in [4].
Fig. 1. (left) A cat-shaped board “naturally” lures players into placing specific pieces at particular locations like the upper or lower-right corners thus making the game history look not uniform in terms of spatial locations; (middle and right) this does not generally happen with standard rectangular boards; however other issues do still occur using these boards like the T-piece placement near the U that gives rise to conflicts between proximity and adherence criteria.

3.4 Evaluation

We evaluated the adaptation capabilities of the Pentomino puzzle system on a standard computer with 2.20 GHz Centrino Pro CPU and 1.96 GB memory under the MS Windows XP operating system. In addition to the usual display, loudspeakers, keyboard and mouse, users were equipped with a microphone to enter spoken commands. The application software is written in JAVA and Prolog. We use a context-free grammar compliant to the Java Speech Grammar Format. We utilize the open-source Sphinx-4 [30] as speech recognition engine and FreeTTS [29] as speech synthesizer. The application dictionary contains approximately 500 words.

In order to test and evaluate the system, we collected the data from 13 participants; seven subjects were female and six were male with ages ranging from 19 to 37. Among the subjects there were 3 native Italian speakers, 2 native English speakers, 6 native German speakers, and 2 native Danish speakers. All subjects reported to be familiar with IT and computer technology in general. Each experiment session lasted approximately 45 minutes plus an additional 30 minutes for both a pre session system explanation and a post session informal talk.

We did not give the subjects any specific task. We just asked them to freely play the game (the one for the average user) to get acquainted with the application and the game rules. Requiring the users to play following a specific sequence of operations would have forced them to adapt their language to fit the assigned task i.e. the very opposite of the goal of our investigation. We provided the users with a digital button that we asked them to push on anytime the spoken command issued was not correctly interpreted by the system as they intended. Each interactive session was logged.

In a second experimental session, we invited the same players one more time. This time though, we asked them to first play the game in guided mode i.e. by following instructions as they appeared on screen. We did so for the system to collect user data and use it to build a model for the playing subject. Each guided session lasted some 45 minutes. We further allocated 45 more minutes for the user to play with the adapted system and 15 minutes for post-session discussion. As in the first experimental session, we equipped the participants with a digital button to signal system interpretation errors.
Participants were not told anything about system adaptation and the goal of our experiment. They were told that they were using another version of the system and that we were testing it against the older one for us to find out which one performs best.

Evaluation occurred by analyzing the log files from both the two interactive sessions and the informal talks that followed each interactive session.

Figures 2 through 4 provide a detailed overview of the performance of both the adaptive system and the more general non-adaptive system for the interpretation of each of the three classes of spatial relationships under consideration.

**Fig. 2.** User adaptation results for the interpretation of topological relations of the kind “near” by 13 subjects; the adaptive system over performs in 5.47% of the cases

**Fig. 3.** User adaptation results for the interpretation of projective relations of the kind “under” by 13 subjects; the adaptive systems over performs in 7.3% of the cases
Table 1 summarizes these results in terms of the average interpretation error and highlights the superior performance of the adaptive system over the more general one. The figures presented can be inferred from Table 2 that provides a tabular representation of the data collected from each of the 13 subjects and over all the spatial relations. The adaptive system achieves a relative error reduction (RER) compared to the more general game system that ranges from 18.97% for topological relations, to 15.5% for projective relations, to finally nearly 10% for the relation “between”. The system created for the average user performs slightly better than the one tailored to the specific needs and habits of the user only in one single case (highlighted in Table 2) and notably for subject 9 and the spatial relation “between”.

When asked about the system that can better interpret user commands, ten users confirmed that it was the second one (i.e. the adaptive one), two users (subject 1 and subject 9) stated that they could not make any final statement, and just one person (subject 8) indicated that the first system (i.e. the non adapted system) was the better one, despite it actually was not.
Table 2. Detailed analysis over all subjects and all spatial relations of the relative error reduction (RER) which is the percentage of increase/decrease of the correct interpretation of the user-adapted system with respect to the general one designated for the average user.

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Spatial Relation</th>
<th>“near”</th>
<th>“under”</th>
<th>“between”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.38%</td>
<td>6.21%</td>
<td>9.21%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>35.25%</td>
<td>9.77%</td>
<td>8.49%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.26%</td>
<td>15.54%</td>
<td>2.91%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>29.34%</td>
<td>11.39%</td>
<td>17.37%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>7.64%</td>
<td>13.96%</td>
<td>16.44%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>26.32%</td>
<td>22.08%</td>
<td>9.38%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15.58%</td>
<td>18.07%</td>
<td>16.53%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>18.16%</td>
<td>4.73%</td>
<td>10.36%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8.24%</td>
<td>22.72%</td>
<td>-3.58%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>22.84%</td>
<td>24.18%</td>
<td>6.84%</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>27.16%</td>
<td>15.44%</td>
<td>5.87%</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>17.87%</td>
<td>22.01%</td>
<td>10.94%</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>22.61%</td>
<td>15.43%</td>
<td>19.46%</td>
<td></td>
</tr>
<tr>
<td>Mean RER</td>
<td>18.97%</td>
<td>15.5%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>9.31%</td>
<td>6.29%</td>
<td>6.39%</td>
<td></td>
</tr>
</tbody>
</table>

It should be noticed that, despite our game system has an operative dialogue manager module that can engage the user in clarification requests to disambiguate incomplete, unclear utterances as well as misrecognition errors from the speech recognizer [5], we disabled such an agent during the evaluation sessions. Our decision is based on the fact that while a working dialogue manager is not necessarily relevant to the analysis of the phenomena we were investigating, it simultaneously does sometimes make wrong decisions thus introducing an additional complexity layer between the system and the user. The absence of a fully functional dialogue manager had the immediate consequence that during evaluation not all user instructions were executed in the desired (or undesired) way by the system. In case of a user instruction that did not directly translate into a game command, the system simply played back a text asking for a repetition of the utterance or an equivalent message.
4 Discussion and Conclusion

Traditionally, user interfaces are built incrementally as an iterative development process that involves the analysis of several factors such as modularity, reliability, efficiency, usability, and goal achievement. The usability aspects require specifically the collection of data about system use in various environments and on different tasks in order to improve the performance of the hypothetic average person representing the final user. The implicit assumption behind this strategy is that the subjects recruited to collect data constitute a representative homogeneous set. Yet, despite such interfaces work fine with the majority of the people in most situations and for a given task, a better system is tailored to the individual user rather than to an abstract typical person.

![Image](image.png)

**Fig. 5.** (1st row, left) Given this game configuration, the dotted area indicates where the user intended to drop the cross-shaped piece for sentence “place the cross between the pieces on-board”; (1st row, right) Expected system response to the user’s sentence; the player however does not issue that command as he anticipates the undesired result based on his mental model of the game; (2nd row, left) instead he changes his strategy and first places the U-shaped piece in the upper left region and then (2nd row, right) he drops the cross next to it (a command semantically equivalent to the originally intended one would have worked as well)

We described an extension of a computational model that approximates human interpretation and judgment of situated language while also accommodating for individual variability and ambiguity in spatial utterances. A proof-of-concept is made by incorporating our approach in the Pentomino board game involving speech user interaction. Observations of the user's behavior provided interaction data that we used to form a model designed to predict future systems actions in response to specific inputs. In the current development, a customized model takes into account playing patterns, level of experience, skill, etc. of the players for the interpretation of spatial expressions that occur while playing. This adaptive approach shows a relevant improvement of the performance with respect to the previous non-adaptive system.
The user model extrapolated by our application during the introductive session should not be confused with the mental model that each user has of the system itself. These two models are intertwined and contribute to the creation of each other. On the one hand, the user model is built by the system during a guided game session and thus is, almost by definition, based on the user’s mental model of the game. On the other one, the opposite is true as well. In fact, we noticed that sometimes players anticipated expected, yet unwanted, system responses to a specific input command by changing the sequence of moves to obtain a desired game configuration (Figure 5). Such behaviors are a mere trick to prevent game moves and do not reflect the user interpretation of specific spatial relationships. Hence, they are potentially damaging to our application since they undermine the collection of consistent training data aimed at reflecting the player’s intentions and game style.

It is worth noticing, that learning based on the user’s interaction behavior may actually misinterpret the user’s intention, when the user is making unexpected mistakes. It is therefore important to assist each user during the guided session to avoid misunderstandings. This fact needs to be considered particularly when an interface is to be used also by people with disabilities or by those who are prone to enter wrong inputs due to their physical limitations.

Due to the way we run the evaluation sessions, our system is little vulnerable to wrong interpretations of user input used as training data. Nonetheless, if an unknown user wants to play the game we need either a preliminary training session to create the user model or to collect training data on the fly while the user is playing. While the former is not always a feasible solution, the latter strategy is however prone to inevitable data interpretation errors. Indeed, erroneous training data is a major problem for any automated learning system, and usually a “smoothing” method is provided to partially account for it. At this point, we do not employ any such method but this is on top of our to-do list.

Our approach can be extended to other domains such as e.g. two-dimensional map based applications, other grid-like scenarios and geographic information systems also in the three dimensional case. A useful augmentation of the current system would be the resolution of spatial relations applied to groups of objects. For instance, an utterance like “select the red pieces under the blue one(s)” requires the selection of a subset of red pieces that are located underneath a specified referent which, in this case, can be itself a subgroup of the set of blue colored pieces. It remains to investigate user variability in the use of referential expressions involving groups of elements.

Acknowledgments: We are grateful to the EU Marie Curie Transfer of Knowledge Programme under Grant #FP6-2002-Mobility-3014491 for partial financial support.

References
