Using HPSG[1] to represent multi-modal grammar in multi-modal dialogue

Crusoe Mao, Tony Tuo, Danjun Liu.
Intel China Research Centre,
Intel China Ltd.
6th Floor, North Office Tower, #06-01 Beijing Kerry Centre
1, GuangHua Road, ChaoYang District,
Beijing 100020, PRC
Email: Crusoe.Mao@intel.com
Tel: (8610)85298800, Fax: (8610)85298717

ABSTRACT
In order to realize their full potential, multi-modal systems need to support not just synchronized integration of multiple input modalities, but also a consistent easy-of-using interface to isolate integration strategies from application ad hoc manner. As the range of multi-modal utterances supported is extended, type of input modalities are increasing, utterances being supported from individual modalities are turning to more complicated, it becomes essential to provide a well-understood and generally applicable common meaning representation for multi-modal utterances. This paper presents a fully formalized declarative statement of multi-modal grammar, the expression we use for the grammar representation draws on unification-based approaches to syntax and semantics, such as head-driven phrase structure grammar (HPSG)[1]. The works presented here show that our approach supports parsing and interpretation of natural human input distributed across the spatial, temporal, and acoustic dimensions. Integration strategies are stated in a high level HPSG based representation supporting rapid prototyping and iterative development of multi-modal systems

KEYWORDS:
Multi-modal Integration, Multi-modality, Head-driven phrase structure grammar (HPSG), Multi-dimensional Parsing

1. INTRODUCTION
Given the complexity of the multi-modal systems that employs visual input spans two spatial dimensions, additional non-spatial acoustic dimension for speech input, and both visual and speech are distributed across the temporal dimension. Unlike the human language, which is naturally discrete and linearly ordered, visual and speech may overlap temporally, and there is no single dimension on which the input is linear and discrete. Besides, multi-modal integration is frequently influenced by underlie communication modalities and the myriad nonlinguistic factors that affect the level of semantic ambiguous and uncertain utterances. So then, what kind of grammar representation that can best model those multi-dimensional, order independent constrains?

HPSG is a constraint-based, lexical list approach to grammatical theory that seeks to model human languages as systems of constraints on typed feature structures. Two virtues in HPSG that we think critical to build multi-modal grammar representation:

- First, HPSG is a constraint-centric type feature structure representation, for factors from communication modalities that could possibly resolve the semantic ambiguous can be expressed in constraint or a cluster of joint constraints
- Second, The meaning of constraint in HPSG is formulated in terms of order-independent, and provides partial grammatical information that can be flexibly consulted in a online integration of heterogeneous types of information. Therefore, integrating multi-dimensional features into a single unified representation is naturally supported.

2. PARSING IN MULTIDIMENSIONAL
Chart parsing methods have proven effective for parsing strings and are commonplace in natural language processing (3). Chart parsing involves population of a triangular matrix of well-formed constituents: CHART(i, j), where i and j are numbered vertices delimiting the start and end of the string. In its most basic formulation, chart parsing can be defined as follows, where * is an operator, which combines
two constituents in accordance with the rules of the grammar

\[ \text{CHART}(I, J) = \text{CHART}(I, K) \times \text{CHART}(K, J). \quad (I < K < J) \]

Crucially, this requires the combining constituents to be discrete and linearly ordered. However, multi-modal input does not meet these requirements: gesture input spans two (or three) spatial dimensions, there is an additional non-spatial acoustic dimension of speech, and both gesture and speech are distributed across the temporal dimension. Unlike words in a string, speech and gesture may overlap temporally, and there is no single dimension on which the input is linear and discrete. So then, how can we parse in this multidimensional space of speech and gesture? What is the rule for chart parsing in multi-dimensional space? Our formulation of multidimensional parsing for multi-modal systems MULTICHART is follow:

\[ \text{MULTICHART}(Z) = \text{MULTICHART}(X) \times \text{MULTICHART}(Y). \quad \text{Where } Y = X + Y, X \land Y = \text{NULL}, X \not\rightarrow \text{NULL}, Y \not\rightarrow \text{NULL} \]

The multidimensional parsing algorithm (Figure 1) runs bottom-up from the input elements, building progressively larger constituents in accordance with the rule set. An agenda is used to store edges to be processed. As a simplifying assumption, rules are assumed to be binary. Each new input received is handled as follows. First, to avoid unnecessary computation, stale edges are removed from the chart. A **timeout** feature indicates the shelf life of an edge within the chart. Second, the interpretations of the new input are treated as terminal edges, placed on the agenda, and combined with edges in the chart in accordance with the algorithm above. Third, complete edges are identified and executed. Unlike the typical case in string parsing, the goal is not to find a single parse covering the whole chart; the chart may contain several complete non-overlapping edges, which can be executed. These are assigned to a category **command** as described in the next section. The complete edges are ranked with respect to probability. These probabilities are a function of the recognition probabilities of the elements, which make up the command. The combination of probabilities is specified using declarative constraints, as described in the next section. The most probable complete edge is executed first, and all edges it intersects with are removed from the chart. The next most probable complete edge remaining is then executed and the procedure continues until there are no complete edges left in the chart. This means that selection of higher probability complete edges eliminates overlapping complete edges of lower probability from the list of edges to be executed. Under certain circumstances, an edge can be used more than once. This capability supports multiple creation of entities. Multiple commands are persistent edges; they are not removed from the chart after they have participated in the formation of an executable command. They are assigned timeouts and are removed when their allotted time runs out. These ‘self-destruct’ timers are zeroed each time another entity is created, allowing creations to chain together.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Figure 1.}
\end{figure}

### 3 Unification-based Multi-modal Grammar Representation

Our grammar representation for multimodal expressions draws on unification-based approaches to syntax and semantics. Multi-modal grammar rules are productions of the form: \( \text{LHS} \rightarrow \text{DTR1} \mid \text{DTR2} \). Where \( \text{LHS} \), \( \text{DTR1} \) and \( \text{DTR2} \) are feature structures of the form indicated above. Following HPSG, these are encoded as feature structure rule schemata. One advantage of this is that rule schemata can be hierarchically ordered, allowing for specific rules to inherit basic constraints from general rule schemata.
The constraints feature indicates an ordered series of functional constraints is common in HPSG. Constraints must be satisfied in order for the rule to apply. Structure sharing in the rule representation is used to impose constraints on the input feature structures, to construct the LHS category, and to instantiate the variables in the constraints. Constraints require certain temporal and spatial relationships to hold between edges. Complex constraints can be formed using the basic logical operators. The constraints are interpreted using a prolog meta-interpreter. This basic backtracking constraint satisfaction strategy is simplistic but adequate for current purposes. It could readily be substituted with a more sophisticated constraint solving strategy allowing for more interaction among constraints, default constraints, optimization among a series of constraints, and so on. The addition of functional constraints is common in HPSG.

4. Conclusion

We have build a multi-modal integrator that can incrementally parse and interpret the HPSG based multi-modal utterance, it has been implemented and will be deployed in our immersive display project and operate in real time, in which, multiple users can collaboratively interact with speech, vision based gesture stylus device and eye gaze. A broad range of multi-modal utterances are supported including combination of speech with multiple gestures from either visual or stylus input and parsing of collections of gestures into complex unimodal commands. We are enhancing this representation to make it sufficiently general to support other input modes and devices including 3D gesture input, mechanic sensors and symbolic gesture input (Sign language), we also are building a multi-modal Chinese input system based on this stated multi-modal integration strategy.

REFERENCES