Clavius: Bi-Directional Parsing for Generic Multimodal Interaction

Frank Rudzicz
Centre for Intelligent Machines
McGill University
Montréal, Canada
frudzi@cim.mcgill.ca

Abstract

We introduce a new multi-threaded parsing algorithm on unification grammars designed specifically for multimodal interaction and noisy environments. By lifting some traditional constraints, namely those related to the ordering of constituents, we overcome several difficulties of other systems in this domain. We also present several criteria used in this model to constrain the search process using dynamically loadable scoring functions. Some early analyses of our implementation are discussed.

1 Introduction

Since the seminal work of Bolt (Bolt, 1980), the methods applied to multimodal interaction (MMI) have diverged towards unreconcilable approaches retrofitted to models not specifically amenable to the problem. For example, the representational differences between neural networks, decision trees, and finite-state machines (Johnston and Bangalore, 2000) have limited the adoption of the results using these models, and the typical reliance on the use of whole unimodal sentences defeats one of the main advantages of MMI - the ability to constrain the search using cross-modal information as early as possible.

CLAVIUS is the result of an effort to combine sensing technologies for several modality types, speech and video-tracked gestures chief among them, within the immersive virtual environment (Boussemart, 2004) shown in Figure 1. Its purpose is to comprehend multimodal phrases such as "*put this* ↘ *here* ↘ .", for pointing gestures ↘, in either command-based or dialogue interaction. CLAVIUS provides a flexible, and trainable new bi-directional parsing algorithm on multi-dimensional input spaces, and produces modality-independent semantic interpretation with a low computational cost.

Figure 1: The target immersive environment.

1.1 Graphical Models and Unification

Unification grammars on typed directed acyclic graphs have been explored previously in MMI, but typically extend existing mechanisms not designed for multi-dimensional input. For example, both (Holzapfel et al., 2004) and (Johnston, 1998) essentially adapt Earley's chart parser by representing edges as sets of references to terminal input elements - unifying these as new edges are added to the agenda. In practice this has led to systems that analyze every possible subset of the input resulting in a combinatorial explosion that balloons further when considering the complexities of cross-sentential phenomena such as anaphora, and the effects of noise and uncertainty on speech and gesture tracking. We will later show the extent to which CLAVIUS reduces the size of the search space.

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 85–90,
Sydney, July 2006.

Directed graphs conveniently represent both syntactic and semantic structure, and all partial parses in CLAVIUS , including terminal-level input, are represented graphically. Few restrictions apply, except that arcs labelled CAT and TIME must exist to represent the grammar category and time spanned by the parse, respectively[1]. Similarly, all grammar rules, $\Gamma_i : LHS \longrightarrow RHS_1\ RHS_2 \ldots RHS_r$, are graphical structures, as exemplified in Figure 2.

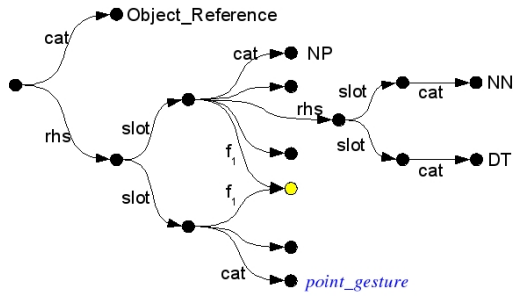

Figure 2: Γ_1 : OBJECT_REFERENCE $\longrightarrow$ NP *click* $\{where(\text{NP} :: f_1) = (click :: f_1)\}$, with NP expanded by Γ_2 : NP $\longrightarrow$ DT NN.

1.2 Multimodal Bi-Directional Parsing

Our parsing strategy combines *bottom-up* and *top-down* approaches, but differs from other approaches to bi-directional chart parsing (Rocio, 1998) in several key respects, discussed below.

1.2.1 Asynchronous Collaborating Threads

A defining characteristic of our approach is that edges are selected *asynchronously* by two concurrent processing threads, rather than serially in a two-stage process. In this way, we can distribute processing across multiple machines, or dynamically alter the priorities given to each thread. Generally, this allows for a more dynamic process where no thread can dominate the other. In typical bi-directional chart parsing the *top-down* component is only activated when the *bottom-up* component has no more legal expansions (Ageno, 2000).

1.2.2 Unordered Constituents

Alhough evidence suggests that deictic gestures overlap *or follow* corresponding spoken pronomials 85-93% of the time (Kettebekov et al, 2002), we must allow for all possible permutations of multi-dimensional input - as in "*put* ↘ *this* ↘ *here.*" vs. "*put this* ↘ *here* ↘ .", for example.

We therefore take the unconvential approach of placing no mandatory ordering constraints on constituents, hence the rule $\Gamma_{abc} : A \longrightarrow$ B C parses the input " C B". We show how we can easily maintain regular temporal ordering in §3.5.

1.2.3 Partial Qualification

Whereas existing bi-directional chart parsers maintain fully-qualified edges by incrementally adding adjacent input words to the agenda, CLAVIUS has the ability to construct parses that instantiate only a subset of their constituents, so Γ_{abc} also parses the input "B", for example. Repercussions are discussed in §3.4 and §4.

2 The Algorithm

CLAVIUS expands parses according to a best-first process where newly expanded edges are ordered according to trainable criteria of multimodal language, as discussed in §3. Figure 3 shows a component breakdown of CLAVIUS 's software architecture. The sections that follow explain the flow of information through this system from sensory input to semantic interpretation.

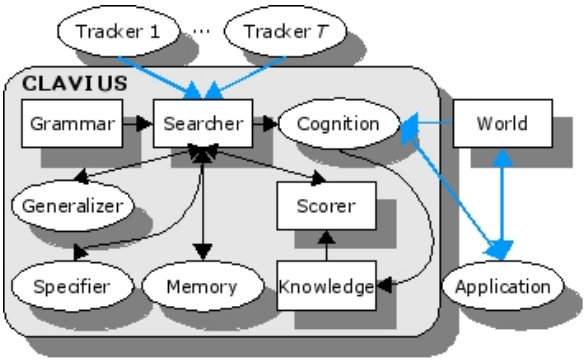

Figure 3: Simplified information flow between fundamental software components.

2.1 Lexica and Preprocessing

Each unique input modality is asynchronously monitored by one of T TRACKERS, each sending an n-best list of lexical hypotheses to CLAVIUS for any activity as soon as it is detected. For example, a gesture tracker (see Figure 4a) parametrizes the gestures *preparation*, *stroke/point*, and *retraction* (McNeill, 1992), with values reflecting spatial positions and velocities of arm motion, whereas

[1]Usually this timespan corresponds to the real-time occurrence of a speech or gestural event, but the actual semantics are left to the application designer

our speech tracker parametrises words with part-of-speech tags, and prior probabilities (see Figure 4b). Although preprocessing is reduced to the identification of lexical tokens, this is more involved than simple lexicon lookup due to the modelling of complex signals.

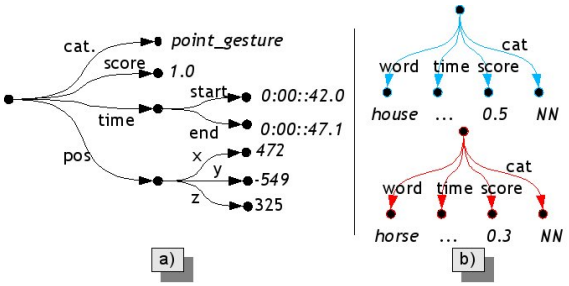

Figure 4: Gestural (a) and spoken (b) 'words'.

2.2 Data Structures

All TRACKERS write their hypotheses directly to the first of three SUBSPACES that partition all partial parses in the search space. The first is the GENERALISER's subspace, $\Xi^{[G]}$, which is monitored by the GENERALISER thread - the first part of the parser. All new parses are first written to $\Xi^{[G]}$ before being moved to the SPECIFIER's active and inactive subspaces, $\Xi^{[SAct]}$, and $\Xi^{[SInact]}$, respectively. Subspaces are optimised for common operations by organising parses by their scores and grammatical categories into depth-balanced search trees having the heap property. The best partial parse in each subspace can therefore be found in $O(1)$ amortised time.

2.3 Generalisation

The GENERALISER monitors the best partial parse, Ψ_g, in $\Xi^{[G]}$, and creates new parses Ψ_i for all grammar rules Γ_i having CATEGORY(Ψ_g) on the right-hand side. Effectively, these new parses are instantiations of the relevant Γ_i, with one constituent unified to Ψ_g. This provides the impetus towards sentence-level parses, as simplified in Algorithm 1 and exemplified in Figure 5. Naturally, if rule Γ_i has more than one constituent ($c > 1$) of type CATEGORY(Ψ_g), then c new parses are created, each with one of these being instantiated.

Since the GENERALISER is activated as soon as input is added to $\Xi^{[G]}$, the process is *interactive* (Tomita, 1985), and therefore incorporates the associated benefits of efficiency. This is contrasted with the all-paths bottom-up strategy in GEMINI (Dowding et al, 1993) that finds all admissable edges of the grammar.

Algorithm 1: Simplified Generalisation

Data: Subspace $\Xi^{[G]}$, grammar Γ
while *data remains in* $\Xi^{[G]}$ **do**
  $\Psi_g :=$ highest scoring graph in $\Xi^{[G]}$
  foreach *rule* Γ_i *s.t.* `Cat` $(\Psi_g) \in RHS(\Gamma_i)$ **do**
    $\Psi_i :=$ `Unify` $(\Gamma_i, [\bullet \rightarrow_{\text{RHS}} \bullet \Rightarrow \Psi_g])$
    if $\exists \Psi_i$ **then**
      `Apply Score` (Ψ_i) to Ψ_i
      `Insert` Ψ_i into $\Xi^{[G]}$
  `Move` Ψ_g into $\Xi^{[SAct]}$

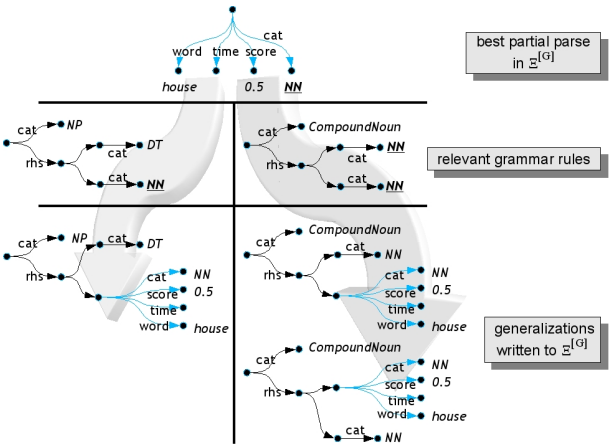

Figure 5: Example of GENERALISATION.

2.4 Specification

The SPECIFIER thread provides the impetus towards complete coverage of the input, as simplified in Algorithm 2 (see Figure 6). It combines parses in its subspaces that have the same top-level grammar expansion but different instantiated constituents. The resulting parse merges the semantics of the two original graphs only if unification succeeds, providing a hard constraint against the combination of incongruous information. The result, Ψ, of specification *must* be written to $\Xi^{[G]}$, otherwise Ψ could never appear on the RHS of another partial parse. We show how associated vulnerabilities are overcome in §3.2 and §3.4.

Specification is commutative and will always provide more information than its constituent graphs if it does not fail, unlike the 'overlay'

method of SMARTKOM (Alexandersson and Becker, 2001), which basically provides a subsumption mechanism over background knowledge.

Algorithm 2: Simplified Specification

Data: Subspaces $\Xi^{[SAct]}$ and $\Xi^{[SInact]}$
while *data remains in* $\Xi^{[SAct]}$ **do**
  $\Psi_s :=$ highest scoring graph in $\Xi^{[SAct]}$
  $\Psi_j :=$ highest scoring graph in $\Xi^{[SInact]}$
  s.t. $\texttt{Cat}\ (\Psi_j) = \texttt{Cat}\ (\Psi_s)$
  while $\exists \Psi_j$ **do**
    $\Psi_i := \texttt{Unify}\ (\Psi_s, \Psi_j)$
    if $\exists \Psi_i$ **then**
      $\texttt{Apply Score}\ (\Psi_i)$ to Ψ_i
      $\texttt{Insert}\ \Psi_i$ into $\Xi^{[G]}$
    $\Psi_j :=$ *next* highest scoring graph from $\Xi^{[SInact]}$ s.t. $\texttt{Cat}\ (\Psi_j) = \texttt{Cat}\ (\Psi_s)$
    ; // Optionally stop after I iterations, for some I
  $\texttt{Move}\ \Psi_s$ into $\Xi^{[SInact]}$

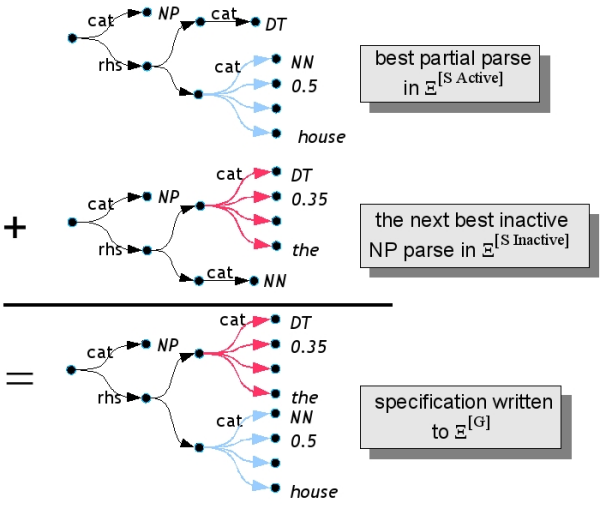

Figure 6: Example of SPECIFICATION.

2.5 Cognition

The COGNITION thread monitors the best sentence-level hypothesis, Ψ_B, in $\Xi^{[SInact]}$, and terminates the search process once Ψ_B has remained unchallenged by new competing parses for some period of time.

Once found, COGNITION communicates Ψ_B to the APPLICATION. Both COGNITION and the APPLICATION read state information from the MySQL WORLD database, as discussed in §3.5, though only the latter can modify it.

3 Applying Domain-Centric Knowledge

Upon being created, all partial parses are assigned a score approximating its likelihood of being part of an accepted multimodal sentence. The score of partial parse Ψ, $\text{SCORE}(\Psi) = \sum_{i=0}^{|S|} \omega_i \kappa_i(\Psi)$, is a weighted linear combination of independent scoring modules (KNOWLEDGE SOURCES). Each module presents a score function $\kappa_i : \Psi \rightarrow \Re_{[0..1]}$ according to a unique criterion of multimodal language, weighted by ω_i, also on $\Re_{[0..1]}$. Some modules provide 'hard constraints' that can outright forbid unification, returning $\kappa_i = -\infty$ in those cases. A subset of the criteria we have explored are outlined below.

3.1 Temporal Alignment (κ_1)

By modelling the timespans of parses as Gaussians, where μ and σ are determined by the midpoint and $\frac{1}{2}$ the distance between the two endpoints, respectively - we can promote parses whose constituents are closely related in time with the *symmetric Kullback-Leibler divergence*, $D_{KL}(\Psi_1, \Psi_2) = \frac{(\sigma_1^2-\sigma_2^2)^2+((\mu_1-\mu_2)(\sigma_1^2+\sigma_2^2))^2}{4\sigma_1^2\sigma_2^2}$. Therefore, κ_1 promotes more locally-structured parses, and co-occuring multimodal utterances.

3.2 Ancestry Constraint (κ_2)

A consequence of accepting n-best lexical hypotheses for each word is that we risk unifying parses that include two competing hypotheses. For example, if our speech TRACKER produces hypotheses "*horse*" and "*house*" for ambiguous input, then κ_2 explicitly prohibits the parse "*the horse and the house*" with flags on lexical content.

3.3 Probabilistic Grammars (κ_3)

We emphasise more common grammatical constructions by augmenting each grammar rule with an associated probability, $P(\Gamma_i)$, and assigning $\kappa_3(\Psi) = P(\text{RULE}(\Psi)) \cdot \prod_{\Psi_c=\text{constituent of } \Psi} \kappa_3(\Psi_c)$ where RULE is the top-level expansion of Ψ.

Probabilities are trainable by maximum likelihood estimation on annotated data. Within the context of CLAVIUS , κ_3 promotes the processing of new input words and shallower parse trees.

3.4 Information Content (κ_4), Coverage (κ_5)

The κ_4 module partially orders parses by preferring those that maximise the joint entropy between the semantic variables of its constituent parses. Furthermore, we use a shifted sigmoid $\kappa_5(\Psi) = \frac{2}{1+e^{-\frac{2}{5}\text{NUMWORDSIN}(\Psi)}} - 1$, to promote parses that maximise the number of 'words' in a parse. These two modules together are vital in choosing fully specified sentences.

3.5 Functional Constraints (κ_6)

Each grammar rule Γ_i can include constraint functions $f : \Psi \rightarrow \Re_{[0,1]}$ parametrised by values in instantiated graphs. For example, the function T_FOLLOWS(Ψ_1, Ψ_2) returns 1 if constituent Ψ_2 follows Ψ_1 in time, and $-\infty$ otherwise, thus maintaining ordering constraints. Functions are dynamically loaded and executed during scoring.

Since functions are embedded directly within parse graphs, their return values can be directly incorporated into those parses, allowing us to utilise data in the WORLD. For example, the function OBJECTAT$(x, y, \&o)$ determines if an object exists at point (x, y), as determined by a pointing gesture, and writes the type of this object, o, to the graph, which can later further constrain the search.

4 Early Results

We have constructed a simple blocks-world experiment where a user can move, colour, create, and delete geometric objects using speech and pointing gestures with 74 grammar rules, 25 grammatical categories, and a 43-word vocabulary. Ten users were recorded interacting with this system, for a combined total of 2.5 hours of speech and gesture data, and 2304 multimodal utterances. Our randomised data collection mechanism was designed to equitably explore the four command types. Test subjects were given no indication as to the types of phrases we expected - but were rather shown a collection of objects and were asked to replicate it, given the four basic types of actions.

Several aspects of the parser have been tested at this stage and are summarised below.

4.1 Accuracy

Table 1 shows three hand-tuned configurations of the module weights ω_i, with $\omega_2 = 0.0$, since κ_2 provides a 'hard constraint' (§3.2).

Figure 7 shows sentence-level precision achieved for each Ω_i on each of the four tasks, where precision is defined as the proportion of correctly executed sentences. These are compared against the CMU Sphinx-4 speech recogniser using the unimodal projection of the multimodal grammar. Here, conjunctive phrases such as *"Put a sphere here and colour it yellow"* are classified according to their first clause.

Presently, correlating the coverage and probabilistic grammar constraints with higher weights ($> 30\%$) appears to provide the best results. Creation and colouring tasks appeared to suffer most due to missing or misunderstood head-noun modifiers (ie., object colour). In these examples, CLAVIUS ranged from a -51.7% to a 62.5% relative error reduction rate over all tasks.

Config	ω_1	$\omega_2^{(*)}$	ω_3	ω_4	ω_5	ω_6
Ω_1	0.4	0.0	0.3	0.1	0.1	0.1
Ω_2	0.2	0.0	0.1	0.3	0.2	0.2
Ω_3	0.1	0.0	0.3	0.3	0.15	0.15

Table 1: Three weight configurations.

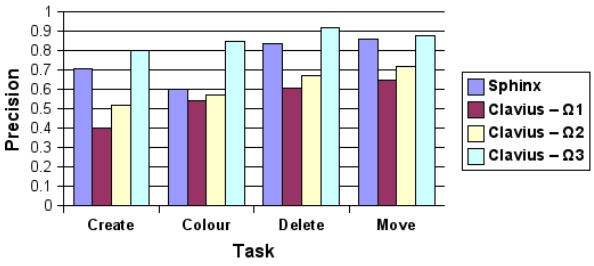

Figure 7: Precision across the test tasks.

4.2 Work Expenditure

To test whether the best-first approach compensates for CLAVIUS ' looser constraints (§1.2), a simple bottom-up multichart parser (§1.1) was constructed and the average number of edges it produces on sentences of varying length was measured. Figure 8 compares this against the average number of edges produced by CLAVIUS on the same data. In particular, although CLAVIUS generally finds the parse it will accept relatively quickly ('CLAVIUS - found'), the COGNITION module will delay its acceptance ('CLAVIUS - accepted') for a time. Further tuning will hopefully reduce this 'waiting period'.

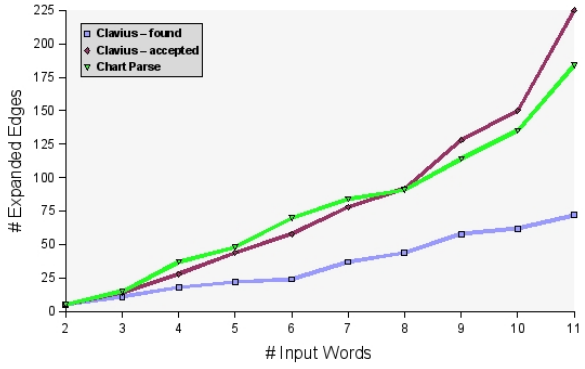

Figure 8: Number of edges expanded, given sentence length.

5 Remarks

CLAVIUS consistently ignores over 92% of dysfluencies (eg. "uh") and significant noise events in tracking, apparently as a result of the partial qualifications discussed in §1.2.3, which is especially relevant in noisy environments. Early unquantified observation also suggests that a result of unordered constituents is that parses incorporating lead words - head nouns, command verbs and pointing gestures in particular - are emphasised and form sentence-level parses early, and are later 'filled in' with function words.

5.1 Ongoing Work

There are at least four avenues open to exploration in the near future. First, applying the parser to directed two-party dialogue will explore context-sensitivity and a more complex grammar. Second, the architecture lends itself to further parallelism - specifically by permitting $P > 1$ concurrent processing units to dynamically decide whether to employ the GENERALISER or SPECIFIER, based on the sizes of shared active subspaces.

We are also currently working on scoring modules that incorporate language modelling (with discriminative training), and prosody-based co-analysis. Finally, we have already begun work on automatic methods to train scoring parameters, including the distribution of ω_i, and module-specific training.

6 Acknowledgements

Funding has been provided by la bourse de maitrisse of the fonds québécois de la recherche sur la nature et les technologies.

References

Ageno, A., Rodriguez, H. 2000 *Extending Bidirectional Chart Parsing with a Stochastic Model*, in Proc. of TSD 2000, Brno, Czech Republic.

Alexandersson, J. and Becker, T. 2001 *Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System* in Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Seattle, WA.

Bolt, R.A. 1980 *"Put-that-there": Voice and gesture at the graphics interface* in Proc. of SIGGRAPH 80 ACM Press, New York, NY.

Boussemart, Y., Rioux, F., Rudzicz, F., Wozniewski, M., Cooperstock, J. 2004 *A Framework for 3D Visualisation and Manipulation in an Immersive Space using an Untethered Bimanual Gestural Interface* in Proc. of VRST 2004 ACM Press, Hong Kong.

Dowding, J. et al. 1993 *Gemini: A Natural Language System For Spoken-Language Understanding* in Meeting of the ACL, ACL, Morristown, NJ.

Holzapfel, H., Nickel, K., Stiefelhagen, R. 2004 *Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures*, in ICMI '04: Proc. of the 6th intl. conference on Multimodal interfaces, ACM Press, New York, NY.

Johnston, M. 1998 *Unification-based multimodal parsing*, in Proc. of the 36th annual meeting of the ACL, ACL, Morristown, NJ.

Johnston, M., Bangalore, S. 2000 *Finite-state multimodal parsing and understanding* in Proc. of the 18th conference on Computational linguistics ACL, Morristown, NJ.

Kettebekov, S., et al. 2002 *Prosody Based Co-analysis of Deictic Gestures and Speech in Weather Narration Broadcast*, in Workshop on Multimodal Resources and Multimodal System Evaluation. (LREC 2002), Las Palmas, Spain.

McNeill, D. 1992 *Hand and mind: What gestures reveal about thought* University of Chicago Press and CSLI Publications, Chicago, IL.

Rocio, V., Lopes, J.G. 1998 *Partial Parsing, Deduction and Tabling* in TAPD 98

Tomita, M. 1985 *An Efficient Context-Free Parsing Algorithm for Natural Languages*, in Proc. Ninth Intl. Joint Conf. on Artificial Intelligence, Los Angeles, CA.