Bootstrap Dialog:
A Conversational English Text Parsing and Generation System

Stephen L. Reed

Texai.org
3008 Oak Crest Ave, Austin TX, 78704
stephenreed@yahoo.com

Abstract

A conversational English text parsing and generation system is described in which its lexicon and construction grammar rules are revised, augmented, and improved via dialog with mentors. Both the parser and generator operate in a cognitively plausible, incremental manner. Construction Grammar is well suited for a precise and robust dialog system due to its emphasis on pairing utterance form with exact logical meaning. Combining lexicon representation and grammar rule representation from the theory of Fluid Construction Grammar, with grammar constructions adopted from Double R Grammar, the system is designed to accommodate wide coverage of the English language.

Introduction

Alan Turing, in his seminal paper on artificial intelligence (Turing 1950), proposed to create a mechanism that simulates a child's mind, and then to subject it to an appropriate course of education thus achieving an artificial general intelligence capable of passing his imitation game. Texai is an open source project to create artificial intelligence. Accordingly, the Texai project is developing conversational agents which are capable of learning concepts and skills by being taught by mentors.

AGI Organization

The Texai AGI, will consist of a vastly distributed set of skilled agents, who are members of mission-oriented agencies that act in concert. Texai agents will be organized as a hierarchical control structure, as described by James Albus (Albus and Meystel 2002). Plans call for users to either converse with a remote, shared Texai instance, or download their own instance for improved performance. Instances host one or more Texai agents and are physically organized as a cloud in which there are no isolated instances. Each Texai agent maintains a cached working set of knowledge safely encrypted, replicated, and persisted in the cloud.

This approach contrasts with Novemente (Goertzel 2006). Although Novemente also employs the artificial child in their development road map, English dialog is not the sole method by which Novemente learns.

One good way to provide mentors for Texai agents is to apply them to human organizations such that each human member has one or more Texai agents as proxies for the various roles the human fills in the organization. The author hopes that Texai will be embraced by, and extend, human organizations. A multitude of volunteers may subsequently mentor the many agents that will comprise Texai.

Bootstrap Dialog

The initial Texai conversational agent is being developed to process a controlled language consisting of a minimal subset of English vocabulary and grammar rules. This controlled language is sufficient to acquire new vocabulary and new grammar rules from its human mentors. The bootstrap dialog system is designed to be taught skills that enhance its programmable capabilities. Texai addresses eight dialog challenges identified by James Allen (Allen et. al. 2000): (1) intuitive natural English input, (2) robustness in the face of misunderstandings, (3) mixed-initiative interaction, (4) user intention recognition, (5) effective grounding and ensuring mutual understanding, (6) topic change tracking, (7) dialog-based response planning to provide the appropriate level of information, and (8) portability so that the dialog system can operate with disparate knowledge domains.

Incremental Processing

The Texai grammar engine operates incrementally, in a cognitively plausible manner. During parsing, words are processed strictly left-to-right with no backtracking. As object referring expressions (e.g. noun phrase) are detected, all semantic and referential interpretations are considered simultaneously for elaboration and pruning by two respective spreading activation mechanisms.

Knowledge Base and Lexicon

The Texai knowledge base is derived from an RDF compatible subset of OpenCyc , and elaborated with RDF extracts from WordNet , Wiktonary, and the CMU Pronouncing Dictionary. The Texai lexicon is available in RDF and N3 format. It features good coverage of English word forms, pronunciations, word senses, glosses, and sample phrases, and an initially modest set of OpenCyc term mappings to word senses.

Knowledge base entities may be mapped into Java objects by the Texai RDF Entity Manager (Reed 2006). Operating in a manner similar to an object-to-relational mapper, the RDF Entity Manager facilitates the automatic retrieval and persistence of Java lexicon and grammar rule objects into the Sesame RDF store.

OpenCyc (Matuszek et al. 2006)
The OpenCyc knowledge base was extracted into RDF format, retaining only atomic terms and contextualized binary assertions. Approximately 130,000 class and individual concept terms are present in the extracted KB. About 12,000 of these terms are linked to WordNet synsets. It is a goal of the Texai project, via dialog with mentors, to complete the mapping of relevant word senses to OpenCyc terms, and to create new Texai terms filling gaps in OpenCyc.

WordNet (Feldman 1999)
WordNet version 2.1 contains lexical and taxonomic information about approximately 113,000 synonym sets. It was fully extracted into RDF for the Texai project.

Wiktonary
This user-authored dictionary is based upon the same platform as Wikipedia. Its XML dump as of September, 2007 was processed into RDF, in a form compatible with WordNet word sense descriptions.

The CMU Pronouncing Dictionary
(http://www.speech.cs.cmu.edu/cgi-bin/cmudict)
This dictionary contains entries for over 125,000 English word forms, and gives each pronunciation as a sequence of phonemes in the ARPABET phonetic alphabet. It is compatible with both the CMU Sphinx automatic speech recognition tools and the CMU Festival speech generation tool. These speech tools are planned as the speech interface for Texai, in addition to its existing text chat interface. The dictionary was processed into RDF, compatible with WordNet word form descriptions.

Merged Texai Lexicon
Using WordNet as the framework, non-conflicting word senses were merged in from Wiktionary. OpenCyc provides corresponding KB terms for 12,000 WordNet synsets. Matching word forms received ARPABET phoneme sequences from the CMU Pronouncing Dictionary.

KB Component	*Nbr. of RDF Statements*
OpenCyc	640,110
WordNet	4,134,543
Wiktonary	3,330,020
The CMU Pronouncing Dictionary	3,772,770
Merged Texai Lexicon	10,407,390

Table 1. KB and Lexicon Components

The Texai KB is physically partitioned by KB component in order for each to fit in main memory (i.e. 2 GB) and provide high performance via a Sesame RDF quad store.

Construction Grammar

Fluid Construction Grammar (FCG) (Steels & De Beule 2006) is a natural language parsing and generation system developed by researchers at emergent-languages.org. The system features a production rule mechanism for both parsing and generation using a reversible grammar. FCG provides a rule application engine in which the working memory (WM) is a coupled semantic and syntactic feature structure. FCG itself does not commit to any particular lexical categories, nor does it commit to any particular organization of construction rules. Like all construction grammars, FCG is a paring between form and meaning. The Texai system extends FCG so that it operates incrementally, word by word, left to right in English. Furthermore, Texai improves the original FCG implementation by adopting a simplified working memory feature structure and by substituting tailored unification for each of the five rule types, instead of using a generic list unification mechanism for construction rule matching. Texai rule formulation also improves on FCG by allowing optional unit constituents in grammar rules, thus reducing dramatically the otherwise large number of explicit permutations.

Double R Grammar

Double R Grammar (DRG) (Ball 2007), previously implemented in the ACT-R cognitive architecture (Ball et al. 2007), is a linguistic theory of the grammatical encoding and integration of referential and relational meaning in English. Its referential and relational constructions facilitate the composition of logical forms. In this work, a set of bi-directional FCG rules are developed that comply with DRG. Among the most important constituents of DRG is the *object referring expression* (ORE), which refers to a new or existing entity in the discourse context. The Texai system maps each ORE to a KB concept. ORE's are related to one another via *situation referring expressions* (SRE), in which the

relation is typically a verb. In the below example, which is formatted in the style of FCG, a PredicatePreposition WM unit is stated to be composed of a Preposition WM unit followed by an ORE construction. Each bi-directional rule consists of units (e.g. ?Prep) having features and attributes. The J unit specifies the head of the rule.

```
(con
 con-PredPrep
 ((category basic-construction))
 ((?Prep
   (category Preposition)
   (referent-subj ?subj)
   (referent-obj ?obj))
  (?ObjReferExpr
   (category ObjectReferringExpression)
   (referent ?obj))
  (?top
   (subunits (== ?Prep ?ObjReferExpr)))
  ((J ?PredPrep)
   (category PredicatePreposition)
   (referent ?obj)
   (referent-subj ?subj))))
```

Figure 1. An Example FCG Rule for a DRG Construction

User Modeling

The knowledge base contains a persistent, contextualized model of each user's belief state. The system is designed to avoid telling the user something that the system knows that the user already knows. This facility is chiefly employed during utterance generation, in which the actual belief state of the user is to be updated with some particular set of propositions.

Discourse Context

The Texai dialog system contains a discourse context for each user interaction session. Each discourse context consists of a list of utterance contexts, each of which represents a single utterance from either the system or the user. Attributes of the utterance context include a timestamp, the speaker identity, the utterance text, a cache of the speaker's preferred word senses, and either the source propositions for a generated utterance, or the understood propositions for a parsed utterance. It is intended that the system learn via reinforcement the number of utterance contexts to maintain, and the degree to which to decay their relative importance.

Incremental Parsing

The dialog system performs incremental utterance parsing in a cognitively plausible manner. As argued by Jerry Ball (Ball 2006) this method avoids possible combinatorial explosions when computing alternative interpretations, and interfaces tightly with automatic speech recognizers. Indeed, it is planned that Texai augment the CMU Sphinx automatic speech recognition tool's language model with respect to scoring alternative recognized words.

Parsing Rule Application

In figure 1 above, Referent variables ?subj and ?obj facilitate the instantiation of logical propositions located throughout a single parsing interpretation. When the above rule is applied in the parsing direction, it matches a Preposition unit in the working memory feature structure being assembled, while binding the ?Prep variable to the corresponding WM unit name. The ObjectReferringExpression WM unit must immediately follow the Preposition in order for this rule to match. As a result of applying the rule in the parsing direction, a new PredicatePreposition WM unit is created in the WM feature structure. This new WM unit has the target Preposition and ObjectReferringExpression WM units as subunits. Incremental processing is facilitated during rule application by hiding already-subordinated WM units, and by focusing rule application on recently created WM units. Incremental processing is achieved by allowing grammar rules to partially match. When processing moves beyond the rightmost required and as-yet unmatched unit of a partially matched rule, its branch of the interpretation tree is pruned.

Kintsch Construction/Integration

Walter Kintsch (Kintsch 1998) proposed a model for reading comprehension, based upon cognitive principles and tested empirically with human subjects. In what he called Construction/Integration, all alternative interpretations are simultaneously considered. For each interpretation, elaborations are constructed in the discourse context (i.e. working memory). Then an iterative spreading activation procedure scores sets of interpretation propositions according to how well connected they are to the concepts initially in the discourse context.

Discourse Elaboration

In the Texai system, discourse elaboration takes place by a marker-passing spreading activation mechanism (Hendler 1998). Discourse elaboration is performed before Kintsch spreading activation so that ambiguous concepts in the input utterance might be inferred to be conceptually related to previously known concepts in the discourse context. In the example presented below, the word “table” is ambiguous. It could either mean cyc:Table, or as part of the multiple word form “on the table“, mean subject to negotiation. There is no known table in the discourse context, but there is a known instance of cyc:RoomInAConstruction. Suppose there exists these commonsense rules in the Texai knowledge base:

- a room may typically contain furniture
- a room may typically have a window
- a room has a ceiling
- a room has a wall

- a room has a floor
- a room has a door
- a room has a means of illumination
- a room can contain a person
- a table is a type of furniture
- a family room is a type of room

Discourse elaboration, via spreading activation, could add furniture, and subsequently table, to the discourse context by activating cached links derived from these rules. A later example will demonstrate.

Pruning By Spreading Activation

Analogous to how automatic speech recognizers operate, the number of retained interpretations in the search space is kept to a specified beam width (e.g. four retained interpretations). At the conclusion of utterance parsing, the highest scoring interpretation is returned as the result.

An Example

The following diagrams were produced during the processing of the example utterance: "the book is on the table". As part of the experiment, the discourse context is primed with knowledge of a room, which is an instance of cyc:RoomInAConstruction, and a book, which is an instance of cyc:BookCopy. Lexical grammar rules matching word stems in the example utterance yield these ambiguous meanings:

- book - a bound book copy
- book - a sheath of paper, e.g. match book
- is - has as an attribute
- is - situation described as
- on - an operational device
- "on the table" - subject to negotiation [a multiword word form]
- on - located on the surface of

These concepts form nodes in a graph, whose links designate a conceptual relationship between two concepts. Marker-passing spreading activation originates at the known discourse terms (e.g. cyc:RoomInAConstruction) and at each significant utterance term (e.g. cyc:Table) and terminates if paths meet (e.g. at cyc:FurniturePiece). When it can be inferred in this fashion that an utterance term and a known discourse term are conceptually related, then that proposition is added to the meaning propositions for subsequent Kintsch spreading activation to resolve ambiguities. The marker-passing spreading activation decays after only a few links to preclude weakly conceptually related results.

The Texai dialog system maintains a tree of parsing interpretation nodes. Each node in the tree is either an input word, such as 'the', or the name of an applied fluid construction grammar rule. Branches in this tree occur when there are alternative interpretations (i.e. meanings) for a word such as "book". The parsing interpretation tree is retained after the parsing process completes so that the user can ask questions about the parsing state (e.g. why a certain grammar rule did not apply as expected).

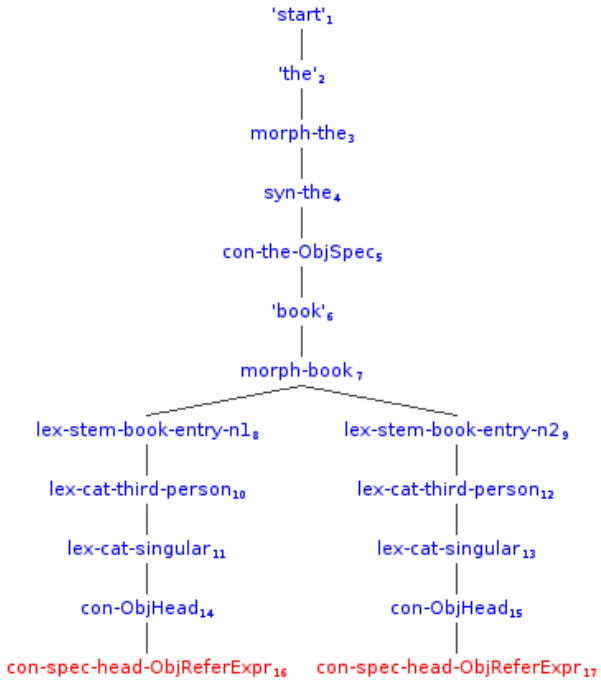

Figure 2. Two alternative interpretations of "book"

Figure 2 depicts the tree of two alternative parsing interpretations for the partial utterance "the book" whose leaves are: node 16 which represents an instance of cyc:BookCopy, and node 17 which represents an instance of texai:SheetsBoundTogetherOnOneEdge. Quoted nodes in the tree represent incrementally parsed words, and the remaining nodes name the applied grammar rule.

According to Walter Kintsch's theory of reading comprehension, spreading activation flows over the nodes of a graph formed by the meaning propositions of the utterance. Links in this graph connect nodes mentioning the same term. The most relevant set of nodes receives the highest activation.

In figure 2 below are the ten propositions from the alternative parsing interpretations of the phrase "the book". In the corresponding figure 3, magenta colored nodes indicate the interpretation: SheetsBoundTogetherOnOneEdge, Cyan colored nodes indicated the alternative interpretation cyc:BookCopy. The yellow nodes indicates prior knowledge - N4 is the prior discourse context knowledge about a cyc:Table, and N5 is the prior discourse context knowledge about a cyc:BookCopy. N1 and N7 are positively connected, which is indicated by a black line, because they share the concept: SheetsBoundTogetherOnOneEdge-2. Node N1 and N10 are negatively connected, which is indicated by a red line, because they represent alternative, conflicting, interpretations.

node	*RDF proposition*
N1	`[texai:SheetsBoundTogetherOnOneEdge-2 texai:fcgStatus texai:SingleObject]`
N2	`[texai:BookCopy-1 rdf:type cyc:BookCopy)]`
N3	`[texai:BookCopy-1 texai:fcgDiscourseRole texai:external]`
N4	`[texai:table-0 rdf:type cyc:Table]`
N5	`[texai:book-0 rdf:type cyc:BookCopy]`
N6	`[texai:BookCopy-1 rdf:type texai:PreviouslyIntroducedThingInThisDiscourse]`
N7	`[texai:SheetsBoundTogetherOnOneEdge-2 rdf:type texai:PreviouslyIntroducedThingInThisDiscourse]`
N8	`[texai:SheetsBoundTogetherOnOneEdge-2 rdf:type texai:SheetsBoundTogetherOnOneEdge]`
N9	`[texai:SheetsBoundTogetherOnOneEdge-2 texai:fcgDiscourseRole texai:external]`
N10	`[texai:BookCopy-1 texai:fcgStatus texai:SingleObject]`

Figure 3. RDF Propositions From Two Alternative Interpretations

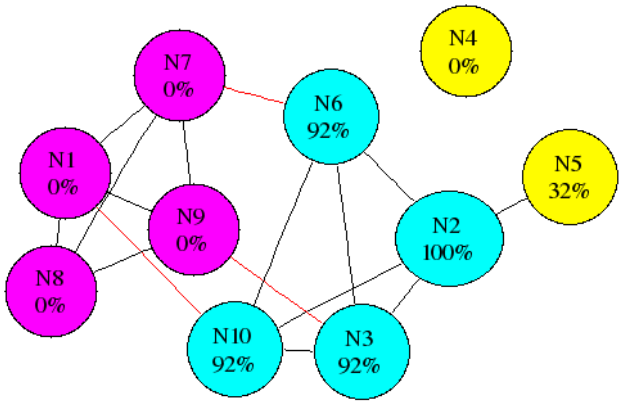

Figure 4. Quiesced Kintsch Spreading Activation Graph

Incremental Generation

The dialog system performs incremental utterance generation. Presently, the dialog planner is rudimentary, and consists of a component that forms a semantic dependence tree from terms in the set of propositions to be communicated to the user. The RDF propositions are gathered by their RDF subject term. One of the terms is heuristically chosen to be the subject of the utterance. Each of the propositions having this term as an RDF subject is selected for the root semantic dependency node. Child nodes are likewise created heuristically for the remaining propositions, grouped by RDF subject term. Incremental generation proceeds in much the same fashion as incremental parsing due to the fact that FCG is bi-directional. As rules match, the resulting utterance is generated left-to-right, word by word. Whenever no rules match, the propositions from the next semantic dependency node are added to the top unit WM feature structure. Pruning of alternative interpretations will be a future research issue. Currently, simple scoring heuristics are:

- prefer fewer words
- prefer to reuse previously uttered words for a given meaning term
- prefer to use words that the recipient is otherwise likely to know

Finally, the resulting generated utterance is trial parsed to ensure that the system can understand what it generates with respect to the discourse context and its model of the user's belief state.

Vocabulary Acquisition

Once the most basic English grammar constructions are hand-coded, it is planed that Texai learn, by being taught, the constructions required for it to comprehend the word sense glosses (i.e. definitions) from WordNet and Wiktionary. By converting this definitional text into crisp, symbolic logic statements, Texai will acquire a degree of commonsense understanding about the defined concepts. The author is postponing grounding most of these concepts in physical perceptions. Initially, the only fully grounded symbols will be those involved with English grammar constructions and vocabulary. That is, Texai will have a grounded perception of what an utterance is, because it can directly sense one and it can generate one. It will likewise have a grounded notion of its mentors. It will be able to sense what each one says to it, and to influence their belief state by what it says.

It might be considered circular and shallow to define the various dictionary words in terms of each other. But this will be sufficient for question answering - e.g. better than an existing search engine on the same corpus. And one might reasonably expect that it will be sufficient for the really important step to soon follow - the acquisition of skills by being taught them.

Grammar Rule Acquisition

At first, rather simple, stylized dialogs will handle idioms and multi-word forms by requesting their corresponding meaning from the user. It remains the subject of future experimentation to find the best way to input logical formulas. The initial idea will be to accept RDF syntax.

After a sufficient vocabulary for describing FCG constructions is developed, then Texai can be taught additional complex grammar rules. This facility will be preceded by a sub-facility that will enable the dialog system to describe to its mentor why a grammar rule succeeded or failed for a particular parsed or generated utterance.

Future Work and Conclusion

Future experiments will explore the hypothesis that bootstrapping works. Here is an example use case:

[user] Do you understand "The book is on neuroscience"?
[texai] Not entirely, neuroscience is not a physical location, how can the book be on it?
[user] The utterance means that the book is about neuroscience.
[texai] Could I generalize "on neuroscience" to any field of scientific discovery, such as "on Astronomy", and mean that the book is about Astronomy?
[user] yes
[texai] Could I generalize field of scientific discovery to some general region, such "on deep space?"
[user] yes, this role can be generalized to any topic.
[texai] Does the utterance "the book is on Mars" entail either the interpretation that (1) the book is physically located on Mars, or (2) the book is about the topic Mars?
[user] yes
…
[user] Do you understand "the book is on fire?"
[texai] Yes, the book is about the topic fire.
[user] There is another, more common, interpretation: There is a fire in which the book is the thing that is burning.
[texai] Could I generalize fire to any decomposition process, such as rot?
[user] no
…
[user] Do you understand "the book is on my to-do list?"
[texai] Yes, the book is about the topic to-do list.
[user] There is another, more common, interpretation: There is a to-do list, and the book is an element of the list.
[texai] I know from Wiktionary that a to-do list is "a list of errands and other tasks - often written on a piece of paper as a memory aid - that one needs or intends to accomplish", but a book is not a task.
[user] metonymy: "the book" means "the task of reading the book"
[texai] Now I understand that "the book is on my to-do list" commonly means that the task of reading the book is a task that I intend to do.

References

Albus, J. S., Meystel, A. M., Intelligent Systems: architecture, design, and control, John Wiley & Sons, New York, USA, 2002

Allen, A., et. al., An Architecture for a Generic Dialogue Shell, In *Natural Language Engineering, Vol. 6*, Cambridge University Press, New York, New York, USA, September 2000.

Ball, J., Heiberg, A. & Silber, R. Toward a Large-Scale Model of Language Comprehension in ACT-R 6. In *Proceedings of the 8th International Conference on Cognitive Modeling*, pages 163-168, 2007

Ball, J. (2006). Can NLP Systems be a Cognitive Black Box? In *Papers from the AAAI Spring Symposium, Technical Report SS-06-02*, pages 1-6, AAAI Press, Menlo Park, California, USA, 2006
http://www.doublertheory.com/NLPBlackBox.pdf

Ball, J. Double R Grammar, 2003
http://www.DoubleRTheory.com/DoubleRGrammar.pdf

Feldbaum, C. ed., WordNet: an electronic lexical database, MIT Press, Cambridge, Massachusetts, USA, 1999

Goertzel, B., The Hidden Pattern: A Patternist Philosophy of Mind, Chapter 15, Brown Walker Press, Boca Raton, Florida, USA, 2006

Hendler, J. A., Integrating Marker-Passing and Problem Solving, Chapter 8 – Cognitive Aspects, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1988

Kintsch, W., Comprehension: a paradigm for cognition, Cambridge University Press, Cambridge, UK, 1998

Matuszek, C., Cabral, J., Witbrock, M., DeOliveira, J. An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, Stanford, CA, USA, March 2006.

Reed, S. L., Semantic Annotation for Persistence, In *Proceedings of the AAAI Workshop on Semantic e-Science*, AAAI Press, Menlo Park, California, USA, 2006

Reiter, E., Dale R., Building Natural Language Generation Systems, Cambridge University Press, Cambridge, UK, 2000

Steels, L. and De Beule, J. (2006) A (very) Brief Introduction to Fluid Construction Grammar. In *Third International Workshop on Scalable Natural Language Understanding* (2006).
http://arti.vub.ac.be/~joachim/acl-ny-06-3.pdf

Turing, A. M. Computing Machinery and Intelligence. In *Mind 59*, 1950.