ChatScript Engine Manual

© **Bruce Wilcox, gowilcox@gmail.com** brilligunderstanding.com

Revision 10/11/2015 cs5.72

This does not cover ChatScript the scripting language. It covers how the internals of the engine work.

Data Structures

Part of understanding CS engine programming is understanding the basic data available to CS. There two basic data structures in CS: the word (dictionary) and the fact. The word represents objects and the fact represents relationships.

Words are the fundamental unit of information in CS. The original words came from WordNet, and then were either reduced or expanded. Word are reduced when some or all meanings of them are removed because they are too difficult to manage. “I”, for example, has a wordnet meaning of the chemical iodine, and because that is so rare in usage and causes major headaches for ChatScript, that definition has been expunged along with some 500 other meanings of words. Additional words have been added, including things that Wordnet doesn't cover like pronouns, prepositions, determiners, and conjunctions. And more recent words like “animatronic” and “beatbox”.

Words have zillions of bits representing language properties of the word (well, maybe not zillions, but 3x64 bytes worth of bits). Many are permanent core properties like it can be a noun, a singular noun, it refers to a unit of time (like “month”), it refers to an animate being, it's a word learning typically in first grade. Other properties result from compiling your script (this word is found in a pattern somewhere in your script). All of these properties could have been represented as facts, but it would have been inefficient in either cpu time or memory to have done so.

Facts are simply triples of words that represent relationships between words. The ontology structure of CS is represented as facts (which allows them to be queried). Words are hierarchically linked using facts (using the “is” verb). Word are conceptually linked using facts with the verb “member”. Word entries have lists of facts that use them as either subject or verb or object so that when you do a query like query(direct_ss dog love ?) CS will retrieve the list of facts that have dog as a subject and consider those.

ChatScript also support variables, again for considerations of efficiency. Variables could have been represented as facts, but it would have increased processing speed, local memory, and user file sizes.

Memory Management

Many programs use malloc and free extensively upon demand. These functions are not particularly fast. And they lead to memory fragmentation, whereupon one might fail a malloc even though overall the space exists. ChatScript follows video game design principles and manages its own memory. It allocates everything in advance and then (with rare exception) it never dynamically allocates memory again, so it cannot fail by calling the OS for memory. And you have control over the allocations upon startup via command line parameters.

This does not mean CS has a perfect memory management system. Merely that it is extremely fast. It is based on mark/release, so it allocates space rapidly, and at the end of the volley, it releases all the space it used back into its own pool.

You might run out of memory allocated to dictionary items while still having memory available for facts. This means you need to rebalance your allocations. But most people never run into these problems unless they are on mobile versions of CS.
The other problem is that memory is not released until the volley is over. So conceivably memory is free but hasn't been freed. But CS supports planning, which means backtracking, which means memory is really not free along the way because the system might revert things back to some earlier state. This problem of free memory mostly shows up in document mode, where reading long paragraphs of text are all considered a single volley and therefore one might run out of memory. CS provides a memory mark and memory free function so you can explicitly control this while reading a document.

Script Execution

The scripting language is heavily dependent upon the prefix character to tell the system how to behave. The script compiler normally forces separate of things into separate tokens to allow fast uniform handling. E.g., "^call(bob hello)" becomes "^call (bob hello)". This predictability allows the system to avoid all the logic involved in knowing where some tokens end and others begin. The other trick the script compiler uses is to put in characters indicating how far something extends. This jump value is used for things like *if* statements to skip over failing segments of the if.

Pos-parsing

The system runs pos-parsing in two passes. The first pass is execution of rule from LIVEDATA/SYSTEM/ENGLISH which help it prune out possible meanings of words. The goal of these rules is to reduce ambiguity without ever throwing out actual possible pos values while reducing incorrect meanings as much as possible. The second pass tries to determine the parse of the sentence, forcing various pos choices as it goes and altering them if it finds it has made a mistake. It uses a "garden path" algorithm. It presumes the words form a sentence, and tries to directly find pos values that make it so in a simple way, changing things if it discovers anomolies.

Queries

Queries like "^query(direct_v ? walk ?) function by having a byte code scripting language stored on the query name "direct_v". This byte code is executed to perform the query.

The Dictionary

The dictionary consists of WORD entries, stored in hash buckets when the system starts up. Once system startup is complete, those buckets are closed and new entries created from user input are all stored in bucket 0, where they can be undo easily. Linear search within this bucket is unimportant because the user won't create many new entries. So when the dictionary performs word lookup, it uses the hash table first, and if it finds nothing there, it uses bucket 0. The hash code is the same for lower and upper case words, but upper case adds 1 to the bucket is stores in. This means all forms of upper case of a word hash the same, so there is only 1 actual print form of an upper case word available.

Marking

The system takes the input and splits it into the original input and a canonical one. Both are "marked". Marking means taking the words of the sentence in order (where they may have pos-specific values)

and noting on each word where they occur in the sentence (they may occur more than once). From specific words the system follows the *member* links to concepts they are members of, and marks those concepts as occuring at that location in the sentence. And concepts may be members of other concepts, and so on up the hierarchy. There exist system functions that allow you, from script, to also mark and unmark words. This allows you to correct or augment meanings.

In addition to marking words, the system generates sequences of 5 contiguous words (phrases), and if it finds them in the dictionary, they too are marked.

Spell Checking

Spell checking takes a word it doesn't recognize and performs a variety of attempts. These in include merging it with adjacent words, splitting it into two words, adding/removing hyphens, or hunting among words whose length is plus or minus one letter, to try making a minimal edit distance.

Script Compiler

In large measure what the compiler does is verify the legality of your script and smooth out the tokens so there is a clean single space between each token. In addition, it inserts “jump” data that allows it to quickly move from one rule to another, and from an “if” test to the start of each branch so if the test fails, it doesn't have to read all the code involved in the failing branch. It also sometimes inserts a character at the start of a patttern element to identify what kind of element it is. E.g., = before a comparison token or * before a word that has wildcard spelling.