Automated Object Frame Creation Based On Large Scale Corpus

James Darren Ratcliff
Department of Computer Science
Midwestern State University
Wichita Falls, TX 76308
falazar@yahoo.com – http://falazar.com/AI

Abstract

The use of frames has long been classified as a simple yet realistic representation of real-world objects and entities and their relationships to each other. One major disadvantage to implementing a frame system is the extensive amount of human time to construct a useful collection of frames. By using statistic contextual data from large corpus of machine-readable text, an automated or semi-automated process can be applied to produce many of these frames.

Introduction

Artificial intelligence's greatest challenge is the overwhelming amount of common sense information that humans use in reasoning. This vast mass of unorganized data must be collected and structured in such a way that it can be used for deductive and other reasoning methods. Frames and ontology are proposed as two major representations for arranging this data. The first, frames, serve as a representation of the object or entity described, while the second describes the hierarchy and other relationships between these items.

These two ideas give a simple organizing structure to the data, converting it into possibly useful information, but do nothing to solve the second issue of the sheer amount of data to be categorized. Expert Systems have the advantage of working with only a small amount of rules and information, which can be listed out by a small number of people in a reasonable amount of time to describe a concentrated field. Doug Lenat of Cycorp created the $25 million 20 year Cyc project with the lofty goal of manually entering in all common sense knowledge known. Currently the Cyc system has over 1,000,000 hand-entered facts, but at a cost of decades, and multiples of that in man-power, and Cyc's goal of common sense understanding is a long way off.

The facts and information of common sense are known to all of us, yet writing them explicitly to a computer would seem to be the lifetime work of many people. The below procedure would augment a Cyc style method with a more efficient technique of data mining, learning these rules and relations from syntactic lexical features found in the large scale corpus. By statistically looking at a significant number of occurrences of terms with their syntax, many relations can be directly gained from this information, and then checked against current knowledge for verification.

Creating the Corpus and Ontology

A large scale text corpus is very easy to create in the age of information and high-speed internet access. For this project there were gathered over a collection of almost 600 recent fiction novels. This untagged machine-readable text corpus provides a large enough base to present many occurrences of most popular terms and phrases used commonly by humans on a daily basis.

A hierarchical ontology is created using the keywords found in the texts, and their definitions from several dictionaries, forming a large IS-A tree similar to the WordNet project from Princeton University. The topological tree given by WordNet was not used because of many discrepancies between the two databases and because of WordNet's choice to often categorize a term by its scientific or very obscure identification as opposed to a more familiar if less precise category. A simple query as to what a "dog" is in WordNet returns domestic dog, canine, carnivore, placental mammal, but none of these gives the simplest answer that a dog is an animal. The dictionary listings give grammatical database for identifying verbs, nouns and other parts of speech, a key prerequisite for the frame creation process. Nouns, and more specifically object nouns: people, places, and things are targeted as major subjects for frames.

N-Gram Statistics and Cooccurrences

Once a keyword or phrase was chosen, the process would begin by identifying the keyword and known information about it, and then collect the context around every occurrence of the keyword. Many statistics and methods were used here, unigram occurrences, bigram and multigram occurrences and semantic word sense resolutions. Unigram and word sense gave back a relative understanding of how popular a term was in common usage, showing that "woman" occurred about 10 times more than "wine", 100 more times than "broom", and 1000 more times than "sea eagle". This could lead to reasoning later that all other things being equal an object the AI is looking at is more likely to be a woman than a sea eagle, if these were the top two choices.

This is a good measure because most other findings of these values are from relatively very small sample sizes from hand-tagged corpora, where two terms may have only been seen one or two times each, leading to a false sense of importance when stating that one term is twice

as popular as another. Because of the large population size of the text used here, all but the more obscure words have at least 100 occurrences.

After the initial context collection is gathered, several parses are made on the data. Bigrams are used to detect word cooccurrences found and to create separate entries in the database for these.

Experimental Results

The more intricate frame building work begins at this stage. The word and its lexical window (words most closely located beside it) are aggregated and processed as well as other top associated words and phrases. Relationships are then built around these findings. An example of a hierarchical relationship between a term and its collocationally related terms is: coffee table, end table, night table, pool table, and trestle table, are all types of table, and would become sub-entities under the table frame entry, thereby enriching the types of tables known in the database.

A few very simple rules can be extracted from the data for each term; by looking up all results of "bed" it can be determined statistically that:

a person can 'lay' or 'lay down' on a bed.
a person can 'sit' or 'sit down' on a bed.
a person can 'get out' of bed.
a person can 'put a person to' bed

It is also known that a bed has a 'side' or 'edge' part which are the same, and this part is often sat on.

With 'closet' some more intricate details are gathered:
a closet has a 'door' Very Sure
a closet is a container, and clothes are often found in a closet. Very Sure, Pretty Sure
a person can 'go to' the closet Very Sure
a person can 'put an OBJECT in' the closet. Very Sure
a person can 'take an OBJECT from' the closet. Pretty Sure
a person can 'hang an OBJECT in' the closet Very Sure

Frame slots such as attributes (color, size, shape), location, near objects, and uses are easily extracted. The second feature in these examples is a Confidence Factor, a number or percentage (not shown) that helps us determine how sure or confident the program is of this fact; this will help with reasoning with confidence needed later on.

Comparing the term closet with other similar terms quickly confirms the ontological definition that a closet is indeed a room, and with reinforcement learning it can be added that a 'Room' frame would also be a container. In this fashion, the ontological IS-A tree can be added to and modified, and frames can be altered by the children of the frames, with a reverse inheritance feature allowing a large amount of children to pass a trait back up to their ontological parent.

Future Work

With the simple data above it is determined that 'clothes' have a very important relationship with closets, and can be conjectured by the software very easily that if someone hangs some object in the closet it is likely to be clothes. This begins to build up a knowledge base of common sense that is lacking in today's software.

The methods above for the most part are very much unsupervised, once a small amount of pattern matching has been applied. A very small amount of effort can return a large payoff because of the similar usage in many constructs of the English language, and with reinforcement learning, and the Confidence Factor, it insures that the standard of high quality of data is retained.

Future methods will most likely incorporate a semi-supervised method in which a human can instruct the AI on certain topics in plain English, but also where the computer may question the human as well regarding missing information or possible connections it has made. This will insure that the computer does not go off on a tangent by using a series of inferences to make poor additions to its knowledge base.

For many words and complex phrases the amount of occurrences is too small, and a doubling to quadrupling of data would increase the information extraction for these, while increasing the confidence levels of the other terms.

Conclusions

The AI software described above has shown that there are ways of extracting usable information from great amounts of plain text data and can be applied to create a large scale reliable common sense knowledge base. Unlike many other projects that require the facts and information be entered manually into a computer, or reducing the knowledge to a narrowly defined topic (Expert Systems), this is a solution that can at least begin to gather the most common knowledge that is inherent to most human beings.

References

Davis R., Shrobe H., and Szolovits P. 1993. What is a Knowledge Representation? *AI Magazine*, 14(1):17-33

Lenat, Douglas B. 2001 Hal's Legacy: 2001's Computer as Dream and Reality.

McCarthy, John; 1959 Programs with Common Sense Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. (Senseval 3) July 25-26, 2004 Barcelona, Spain.

Schubert, Lenhart; 2002 Can We Derive General World Knowledge from Texts.

Singh, P; Lin, T; Mueller, E; Lim, G; Perkins, T; and Zhu, W 2002. Open Mind Common Sense: Knowledge acquisition from the general public. *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*.

Sowa, John F., 1999 *Knowledge Representation: Logical, Philosophical, and Computational Foundations.*