

SaskNet: Automatically Creating a Large Scale Semantic Knowledge Network

Brian Harrington

Computing Laboratory, Oxford University

brian.harrington@comlab.ox.ac.uk

Semantic knowledge resources such as WordNet^[1] and the Cyc project^[2] have a great many potential uses in fields ranging from Artificial Intelligence and Machine Translation to Web Search and Question Answering. Unfortunately the wide breadth of coverage these tasks require make building resources very costly and labour intensive. Traditionally these resources would require decades of work by large and highly specialized teams to generate a large enough knowledge base to be useful for most applications.

The SaskNet (Spreading Activation based Semantic Knowledge Network) project is an attempt to automatically generate a large scale semantic knowledge network using NLP techniques. If successful, SaskNet would not only save a great deal of time and effort spent in developing semantic knowledge resources, but could also potentially generate a resource on a much larger scale than has ever been possible.

SaskNet translates sentences of English text (in this case taken from the Wall Street Journal) and processes each sentence into a semantic network fragment. For example, in processing the sentence “IBM bought Lotus” two object nodes would be created to represent IBM and Lotus respectively, and then a directed arc labelled “bought” would be created travelling from the IBM node to the Lotus node.

Once a network fragment has been created for a sentence, a spreading activation based algorithm is used to determine which nodes, if any, in the fragment are semantically identical to nodes in the existing knowledge network. For example, if our knowledge network already had a node with the label “Lotus Development Corporation”, which contained links to other information in the network, the algorithm should map our “lotus” node to this “Lotus Development Corporation” node rather than simply inserting a new node into the network.

Mapping two nodes which refer to the same entity together (Object Resolution) is a very difficult task. Two sentences may refer to an identical object using entirely different terms, or may use the same word to label two completely different objects (as we may run into if our hypothetical network already had an entry for “lotus flowers”). SaskNet attempts to solve this problem by resolving objects using semantic information rather than just syntactic.

In order to decide which nodes should be mapped together, SaskNet uses an algorithm based on spreading activation. Similar to the neural network paradigm, spreading activation allows SaskNet to “fire” a node within the network, which

then causes activation to spread from that node along all of its connected links to its neighbouring nodes. Once a node receives a certain amount of activation, it then fires and sends out more activation. This process allows SaskNet to determine the strength of all connections between two nodes in the network (which we take as a measure of their semantic closeness).

Spreading activation allows SaskNet to determine which objects should be mapped together using not only their syntactic similarity (i.e., String similarity, part of speech similarity, etc), but also their syntactic similarity (i.e., the objects and concepts to which they are related).

Large scale semantic resources are very useful in a wide variety of tasks, however manually creating these resources is cost prohibitive. The SaskNet project hopes to automatically create a semantic network that will not only save time and cost, but will also also create a network large enough to be useful to projects which can not use currently available resources.

[1] *Christiane Fellbaum, editor. WordNet : An Electronic Lexical Database. MIT Press, Cambridge, Mass, USA, 1998.*

[2] *Douglas B. Lenat. Cyc: A Large-scale Investment in Knowledge Infrastructure. Communications of the ACM, 38(11):33 – 38, 1995.*