ASKNet: Automatically Generating Semantic Knowledge Networks

Brian Harrington
Oxford University Computing Laboratory
Wolfson Building, Parks Rd, Oxford. OX1 3QD
+44 (0)1865 273870
http://www.brianharrington.net/asknet
brian.harrington@comlab.ox.ac.uk

Abstract

The ASKNet project uses a combination of NLP tools and spreading activation to transform natural language text into semantic knowledge networks. Network fragments are generated from input sentences using a parser and semantic analyser, then these fragments are combined using spreading activation based algorithms.

The ultimate goal of the project is to create a semantic resource on a scale that has never before been possible. We have already managed to create networks more than twice as large as any comparable resource(1.5 million nodes, 3.5 million edges) in less than 3 days. This report provides a summary of the project and its current state of development.

Introduction

The goal of the ASKNet (Automated Semantic Knowledge Network) project is to develop a system which can automatically extract knowledge from natural language text, and use a combination of existing NLP tools and spreading activation theory to build a large scale semantic network to represent that knowledge.

ASKNet translates natural language sentences into fragments of a semantic network. Each of these fragments is taken as an update to the existing knowledge network, and so the fragment (and thus the sentence) is interpreted within the reference frame provided by all the other sentences already processed. This means that the network improves with each sentence it processes, and knowledge gained from previous documents can help with future processing.

Motivation

Semantic networks are a valuable part of many research projects in Artificial Intelligence. For this reason, projects such as ConceptNet (Liu & Singh 2004) and the Cyc Project (Lenat 1995) have spent years manually constructing networks. However, manual construction severely limits the coverage and scale that a semantic resource can possibly achieve. After more than a decade of work, the largest semantic networks available have on the order of 1.5-2.5 million relations connecting 200,000 - 300,000 nodes (Matuszek *et al.* 2006). These networks have had some success in many tasks such as question answering (Curtis, Matthews, & Baxter 2005) and predictive text entry (Stocky, Faaborg, & Lieberman 2004). However, many tasks either require a domain specific knowledge base to be created quickly, or require much wider coverage than is possible to achieve in manually created networks. Automatic generation fulfils both of these requirements.

The most successful automatic generation system to date is MindNet (Dolan *et al.* 1993). Started in 1993 at Microsoft Research, MindNet uses a wide coverage parser to extract pre defined relations from dictionary definitions. To illustrate the difference in creation time for automated construction over manual creation, the MindNet network of over 150,000 nodes connected by over 700,000 relations (roughly half the size of the ConceptNet or Cyc networks) can be created in a matter of hours on a standard personal computer (Richardson *et al.* 1998).

MindNet is a step forward in the effort to automatically create semantic resources. However, we feel that this project limits itself greatly by constraining its relations to a small, pre-defined set, and using a methodology which is highly dependant on well formed data such as dictionaries, and is not extensible to more generalised text.

The ASKNet project hopes to improve upon existing systems by extracting the relations from the text itself, using a wide coverage parser trained on newspaper text and also by allowing arbitrarily complex node structures to be linked together in the same manner as individual nodes. This means that ASKNet can accommodate a much wider variety of information, use more varied sources of input, and extract more information than any other system currently in development.

The Network

ASKNet uses the Clark & Curran Parser (Clark & Curran 2004) and the semantic analysis program Boxer (Bos 2005) to transform input sentences into Discourse Representation Structures (DRS) (Kamp & Reyle 1993). In DRS format, the constituent objects of each sentence and the relationships between these objects are identified and output in the form of first order logic predicates.

Once the semantic analysis of a sentence has been completed, ASKNet uses the DRS to create a semantic network fragment from each sentence. The network created is a di-

Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

rected, nested graph with nodes representing objects and concepts and edges representing semantic relations. A simplified network is shown in Figure 1.

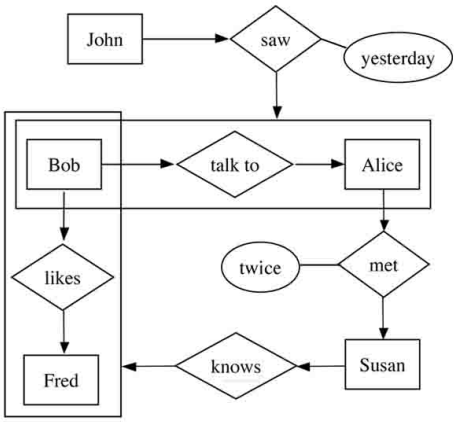

Figure 1: A simplified Semantic Network created from the sentences "John saw Bob talk to Alice yesterday. Alice met Susan twice. Susan knows Bob likes Fred."

One important feature of the ASKNet network is that the relations are automatically parsed from the input text instead of being taken from a set of pre-defined relations, as is the case in most other resources of this type. This vastly increases the expressiveness, and thus the overall power, of the network.

Another important feature of the network is its nested structure. ASKNet allows nodes and relations to be combined to form complex nodes which can represent larger and more abstract concepts. These complex nodes can be combined with further relations to represent even more complex concepts. This unbounded nesting allows ASKNet to express a very wide variety of information without imposing a rigid hierarchical structure.

All of the relations in the network have a weight which represents the confidence of the network in the "real world" existence of the relation and also its salience. Factors such as the confidence of the initial sentence parse, the source of the information, how recently the information has been encountered and the number of different sources verifying the information could all affect the weight given to a relation, depending on the desired application. For example, the current version of the program sets the weights of the relations based on the number of times that relation has been seen, and also increases the weights of relations gathered from headlines over those gathered from the body of a document.

Spreading Activation

Spreading activation mimics the way neural activity spreads in the human brain. When a node in a semantic network receives a certain amount of activation, it fires, sending activation to all neighbouring nodes. By firing a node and monitoring the spread of activation, ASKNet can determine which nodes are semantically related in the network.

ASKNet uses spreading activation to decide which object nodes should be mapped together in the network. In a process known as Object Resolution, two nodes are fired and their activation spread is monitored. If two nodes have very similar patterns of spreading activation, they likely refer to the same real world entity, and therefore should be collapsed into a single node.

Object resolution is an important part of ASKNet as it allows information from multiple data sources to be combined into a single cohesive network.

Current State of Development

The ASKNet project is currently still in its testing phase, however the early results are very promising. In recent tests, the ASKNet system was able to build a semantic network of over 1.5 Million nodes and 3.5 Million edges (more than twice as large as any existing network) in less than 3 days.

The next stage of the project is the development of an appropriate evaluation metric. This will allow us to better finetune the parameters of the spreading activation algorithms, while also providing us with an overall assessment of the work which has already been completed.

For more information, visit the ASKNet project website at: http://www.brianharrington.net/asknet.

References

Bos, J. 2005. Towards wide-coverage semantic interpretation. In *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, 4253.

Clark, S., and Curran, J. R. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, 104–111.

Curtis, J.; Matthews, G.; and Baxter, D. 2005. On the effective use of Cyc in a question answering system. In *Papers from the IJCAI Workshop on Knowledge and Reasoning for Answering Questions*.

Dolan, W. B.; Vanderwende, L.; ; and Richardson, S. 1993. Automatically deriving structured knowledge base from on-line dictionaries. In *Proceedings of the Pacific Association for Computational Linguistics*.

Kamp, H., and Reyle, U. 1993. From discourse to logic : Introduction to modeltheoretic semantics of natural language. *Formal Logic and Discourse Representation Theory*.

Lenat, D. B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33 – 38.

Liu, H., and Singh, P. 2004. ConceptNet a practical commonsense reasoning tool-kit. *BT Technology Journal* 22:211 – 226.

Matuszek, C.; Cabral, J.; Witbrock, M.; and DeOliveira, J. 2006. An introduction to the syntax and content of Cyc. In *2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*.

Richardson, S. D.; Dolan, W. B.; ; and Vanderwende, L. 1998. MindNet: Acquiring and structuring semantic information from text. In *Proceedings of COLING '98*.

Stocky, T.; Faaborg, A.; and Lieberman, H. 2004. A commonsense approach to predictive text entry. In *Proceedings of Conference on Human Factors in Computing Systems*.