ASKNet: Automated Semantic Knowledge Network

BRIAN HARRINGTON AND STEPHEN CLARK

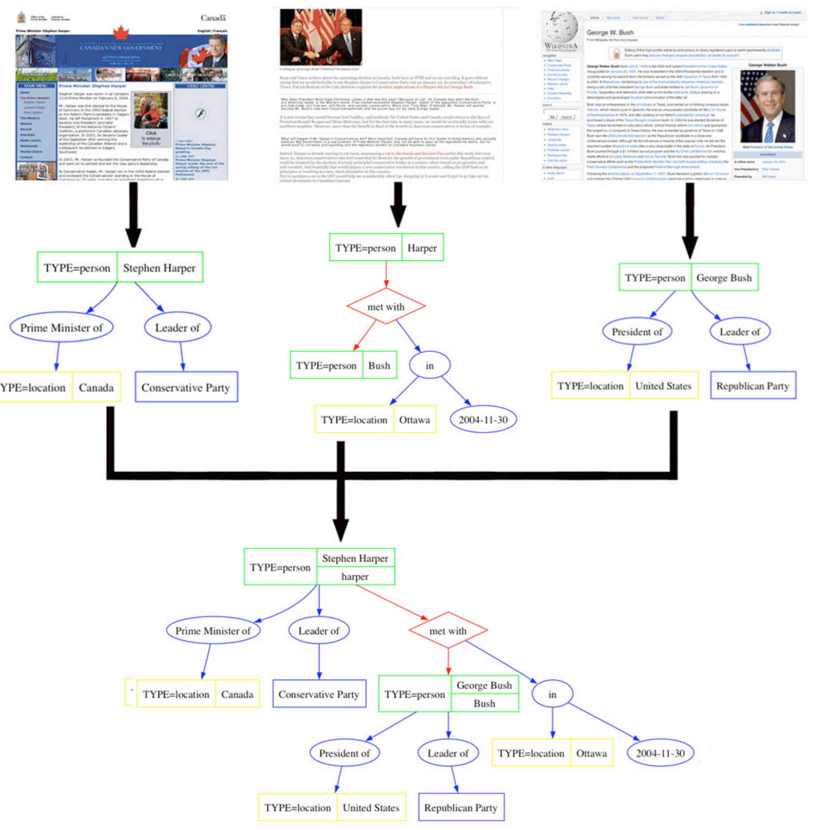
Oxford University Computing Laboratory
Wolfson Building, Oxford, OX1 3QD,UK

ABSTRACT

The ASKNet system is an attempt to automatically generate large scale semantic knowledge networks from natural language text. State-of-the-art language processing tools, including parsers and semantic analysers, are used to turn input sentences into fragments of semantic network. These network fragments are combined using spreading activation-based algorithms which utilise both lexical and semantic information. The emphasis of the system is on wide-coverage and speed of construction. In this paper we show how a network consisting of over 1.5 million nodes and 3.5 million edges, more than twice as large as any network currently available, can be created in less than 3 days. We believe that the methods proposed here will enable the construction of semantic networks on a scale never seen before, and in doing so reduce the knowledge acquisition bottleneck for AI.

What is ASKNet

The ASKNet (Automated Semantic Knowledge Network) [1] system automatically extracts knowledge from natural language text and, using a combination of Natural Language Processing (NLP) tools and spreading activation theory, builds a semantic network to represent that knowledge.

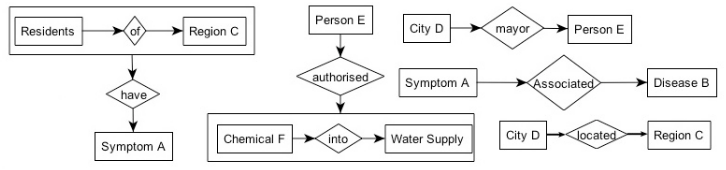


ASKNet uses the **Clark & Curran Parser** [2] and the semantic analysis tool **Boxer** [3] in order to extract relations directly from the text, and therefore, unlike typical resources of its kind, ASKNet does not limit the set of possible relations it can extract.

Motivation

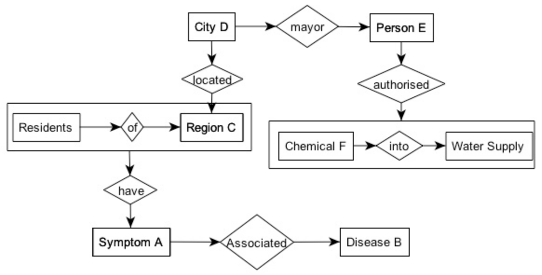
Automatically constructing semantic networks is an important AI task, but it also has immediate practical benefits. Years of work has been put into manually creating semantic networks on a scale of 1.5 - 2.5 million relations connecting 200,000 - 300,000 nodes [4]. By automatically constructing networks, we can build resources of similar or larger sizes in a matter of days. Automatic construction also makes it possible to create networks with much wider coverage than is possible to achieve in manually created networks.

Information Integration

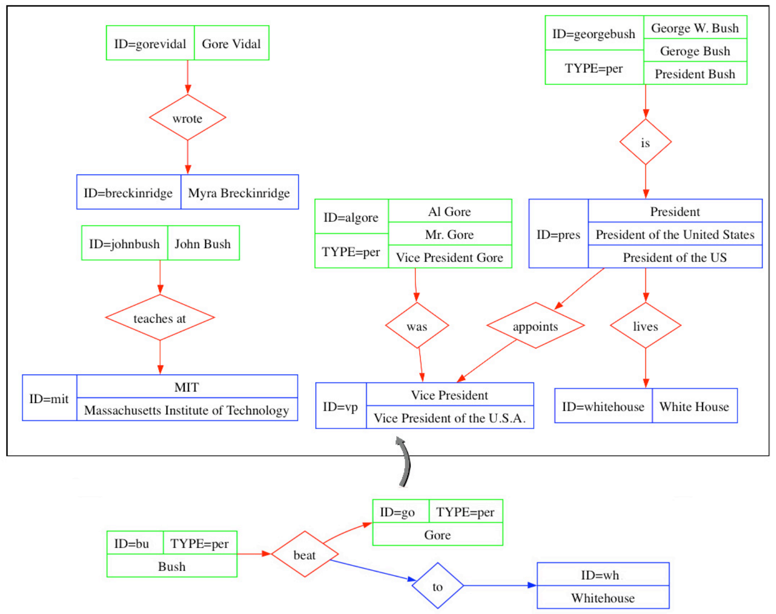


One of the most important features of ASKNet is its ability to combine information from multiple sources into a single cohesive network. Mapping co-referent nodes together provides a great deal of the potential power of the network, and transforms it from a series of disconnected fragments into a single connected resource.

Information Integration allows ASKNet to discover connections between entities which have never been referenced in the same document. By integrating the fragments of information into a cohesive network we can easily determine the relationship between these entities.



The Update Algorithm



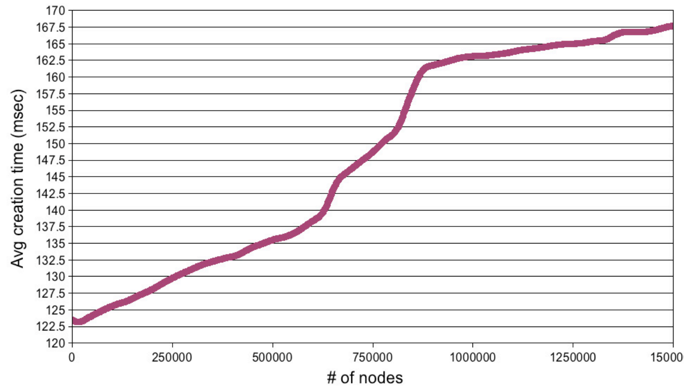
ASKNet uses a technique known as **Spreading Activation** whereby firing a node sends activation out to all of the neighbouring nodes. By firing one or more of the nodes, and analysing the pattern of activation spread in the network, we can determine the strength of the connections between various entities.

In the example above, we can determine where to map the "bu", "go" and "wh" nodes by firing all nodes which have similar syntactic properties (for example firing both the "algore" and the "gorevidal" node to correspond to the "go" node) and analysing their activation patterns. Firing the "georgebush" and "algore" nodes would cause activation feedback, while firing the "gorevidal" and "johnbush" nodes would not. Thus we can determine that it was George Bush who beat Al Gore to the White House, and not another of the possible semantic combinations.

Speed & Network Size

Total Number of Nodes	1,500,413
Total Number of Edges	3,781,088
Time: Parsing	31hrs : 30 min
Time: Semantic Analysis	16 hrs: 54 min
Time: Building Network & Information Integration	22 hrs : 24 min
Time: Total	70 hrs : 48 min

By processing approximately 2 million sentences of newspaper text, ASKNet was able to build a network of over 1.5 million nodes connected by over 3.5 million links in less than 3 days. This is a vast improvement over manually created networks, which have taken years to create networks of less than half this size.[4].



As the size of the network increases, the time required to add additional nodes begins to increase exponentially. However, the localised nature of spreading activation ensures that after the network reaches a threshold size (in this case approximately 850,000 nodes) the portion of the network affected by any given firing ceases to grow and thus the time needed to expand the network becomes linear with respect to the number of nodes added.

Discovering Novel Facts via Connectivity Ranking

ASKNet	Correlational
Bill Clinton	Bill Clinton
Hillary Rodham Clinton	Kenneth Star
Richard Socarides	Monica Lewinsky
Moica Lewinsky	Hillary Clinton
Alfred P. Murrah	Al Gore
Al Gore	Linda Tripp
Kenneth Star	Paul Hosefros
Mary Nell Lehnhard	Henry Hyde
...	...

ASKNet can be used to rank the strength of the connections between pairs of entities. By comparing this ranking with a simple correlational ranking obtained by co-occurrence of entities within documents, we can determine which relationships are "novel" (shown above in red) and can not be found simply by reading individual documents, and which relations are "spurious" (shown above in blue) and do not denote real-world relationships.

References

[1] Harrington B. and Clark S. 2007 ASKNet: Automated Semantic Knowledge Network. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*

[2] Clark, S., and Curran, J. 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)*

[3] Bos, J.; Clark, S.; Steedman, M.; Curran, J. R.; and Hockenmaier, J. 2004. Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*,

[4] 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*.