Discovering Novel Biomedical Relations using ASKNet Semantic Networks

Brian Harrington
Medical Informatics Group
University of Wisconsin-Milwaukee
2400 E Hartford Ave
Milwaukee WI, USA
brian@brianharrington.net

ABSTRACT

One of the greatest challenges facing the biomedical community at the moment is information overload. It is simply not feasible for researchers to read and absorb the sheer quantity of information available. For this reason, many important relationships are going undiscovered. This paper details the ASKNet project, and explains how it could be used to develop a tool that would allow biomedical researchers to firstly identify relationships of interest between entities such as proteins and genes, and secondly to automatically decide which of these relationships is genuinely novel in nature. In essence, ASKNet is capable of being the first fully autonomous system to produce genuine scientific discoveries.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods Semantic networks

General Terms

Algorithms

Keywords

Semantic Networks, Spreading Activation, Relation Discovery

1. INTRODUCTION

The application of natural language processing (NLP) techniques to biomedical data has proven to be very beneficial to researchers in helping them mine information from text, and filter that information into useful data. However, the ability to combine information to produce new insights has eluded the computer science community, leaving the discovery of truly novel information solely the domain of humans. However, the sheer volume of information available to research scientists, particularly in fields such as biomedicine, precludes the possibility of a single individual reading more than a fraction of the published material available to them. Thus, the information bottleneck undoubtedly results in important connections not being made, simply because no single individual ever happens to read the combination of documents necessary to make the connection apparent.

One of the goals of the ASKNet project, is to provide a means for researchers to discover relationships between entities that would not otherwise be possible, and eventually, to produce a system which can autonomously discover truly novel information. This research has implications in a number of fields, but in no area of research could this system be of more benefit than bioinformatics. The ability for researchers to process the tremendous volume of information available to them and automatically discover novel relationships between genes, proteins, diseases and environmental factors, among others, could completely revolutionize the way biomedical research is performed.

2. ASKNET

ASKNet is a system for automatically creating semantic knowledge networks from natural language texts. The ASKNet networks are psycholinguistically inspired, and heavily based on spreading activation theory[4]. The main focus of the ASKNet project has been to create a biologically inspired system for the collection, integration and management of textual data. In particular, ASKNet focus on using spreading activation algorithms, based upon the working of the human brain, in order to integrate information from multiple sources and create a single cohesive information resource.

2.1 Semantic Networks

The representation of biomedical data in Semantic Networks is well established by projects such as the Unified Medical Language System (UMLS) project[1]. However, most of these networks are manually created and managed, and thus require a great deal of both time and resources to construct. Furthermore, they cannot keep up with newly published information, which can often be the most valuable for researchers.

The ASKNet semantic network formalism is based on an entity relationship graph, with nesting structures to allow for complex concepts to be built from simple relations. A simple example is given in Figure 1, showing how relations can connect simple atomic entities (e.g., ABC Inc and Susan), or complex concepts built from simpler relationships (e.g.,

Bob and the concept of ABC Inc moving to London). This nested structure allows for arbitrarily complex concepts to be built and linked together, resulting in a very powerful and expressive formalism. All of the entities, relations and attributes in the networks are taken directly from text, and whereas some existing approaches train on a small set of relations, thus limiting the expressiveness of their network, ASKNet relations are limited only by the language used in the original text.

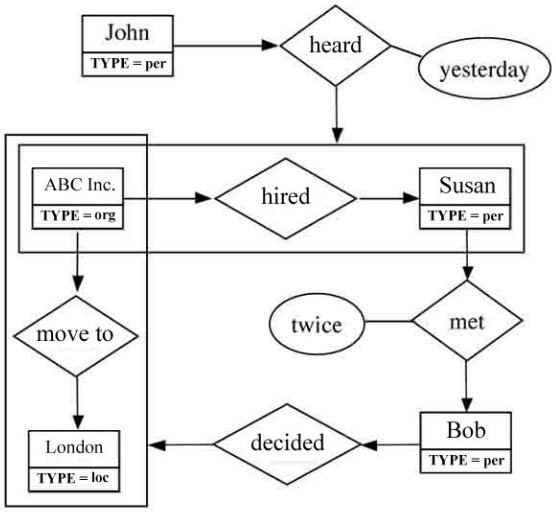

Figure 1: A simplified ASKNet network created from the sentences "Yesterday John heard that ABC Inc. hired Susan. Bob decided that ABC Inc. will move to London. Susan met Bob twice."

All relations and attributes in an ASKNet network are assigned a *strength*, which can be determined by a number of factors, but essentially represents the certainty and the relevance of a particular relation[5]. These links determine the amount of activation that is sent along the links during the spreading activation algorithms.

2.2 Information Integration

There are many systems in existence that can mine information from documents, and these have been of great use to a wide variety of scientific communities. However, there are some types of information that can only be gained by combining the data from multiple sources. ASKNet uses spreading activation algorithms, similar to those found in the human brain, in order to map together information found in multiple sources and create a single unified knowledge resource.

The spreading activation algorithms used by ASKNet are based on similar algorithms used in neural networks. Any entity in a network can be given an amount of *activation*; if the activation of a node exceeds the node's firing threshold, the node will fire, sending its activation to its neighbours with the amount of activation sent to each neighbour being determined by the relative strength of the connecting relation. ASKNet can then analyze the manner in which the activation spreads through the system in order to determine the overall degree of relatedness of nodes.

After processing text with a dependency parser[3] and semantic analyzer[2], ASKNet produces network fragments representing the information in a single document. Each fragment is then put through the *update algorithm*[7] which uses spreading activation to decide which nodes in the fragment map onto existing nodes in ASKNet's global knowledge network. Through this process, new information is "learned" and combined with existing knowledge to produce one single unified knowledge network.

An example of the benefits of information integration is shown in Figure 2. In the first network, we have a series of discrete network fragments, which may be able to provide information about specific facts, and could be useful for data mining. However, in the second network, the fragments have been integrated into a single cohesive network. It is only when this network is integrated that we can establish a pathway between *Chemical F* and *Disease B*. Thus, after the information integration step, we have produced a semantic network which is more than the sum of its parts.

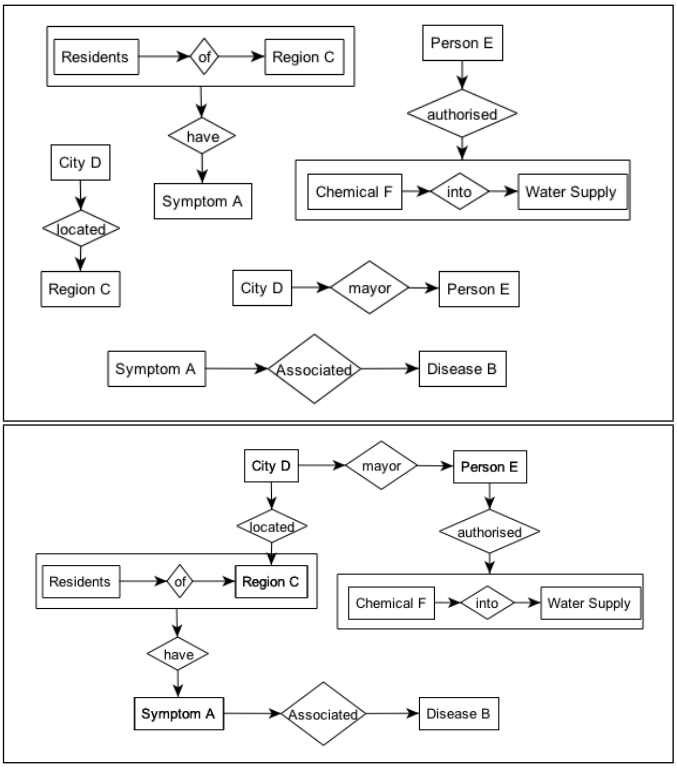

Figure 2: (*top*) A series of network fragments taken from disparate sources. (*bottom*) The network fragments after having been integrated into a single network.

2.3 Scale & Quality

Producing large scale integrated networks is of little benefit, if those networks cannot be built efficiently and to a high standard of quality. Manually created networks will likely be of a higher quality due to the imperfect nature of natural language processing, but in many cases, the benefits

gained by being able to produce large networks quickly far outweighs the decrease in precision.

In order to establish the speed with which networks can be created, we have produced a network based on text mined from Wikipedia with 32 million relations connecting over 2.1 million nodes; roughly twice the size of the network produced by the UMLS project[1], in just under 4.5 days, including all data parsing, semantic analysis and integration[10].

It is important to note that, while the spreading activation algorithms are exponential in nature, they are also localized. This means that as the size of the network grows, the average time to integrate a single node grows exponentially, but only until the network reaches a critical size, where the localized nature of the spreading activation means that the new nodes added to the network do not affect the algorithms. As can be seen in Figure 3, when the network is first being constructed, the exponential nature of the algorithms causes the time to increase at a rapid rate, but as the network grows, the average time to create a node levels off, and grows linearly with the size of the network[8], thus allowing for efficient large scale network creation.

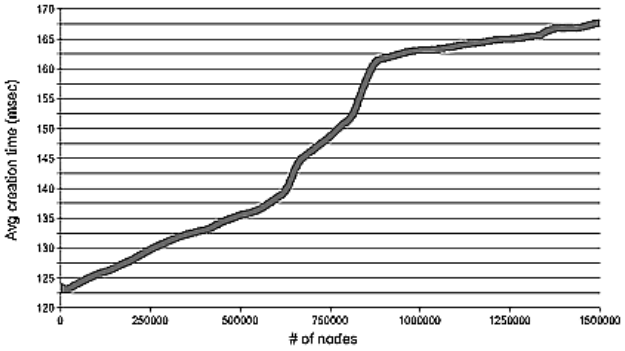

Figure 3: A graph showing the average node creation time vs. the total number of nodes in a network.

Efficient creation is only beneficial as long as the resulting networks are of high enough quality to be useful in the scientific community. For this reason, ASKNet networks must be evaluated both directly, by human judges, and indirectly by using the networks in a standard test.

A manual evaluation was carried out on networks built from the 2006 Document Understanding Conference documents. Human evaluators were asked to judge network "cores", extracted from the full scale networks, and for each connected pair of nodes, give a mark indicating whether all paths connecting them were correct. If there were any mistakes in any connecting paths, or if any nodes were not appropriately integrated, the entire pair was marked as incorrect. In total 5 networks were evaluated by 3 human judges, obtaining a final precision score of 79.1% [8]. The full results are provided in Table 1, and an example network core is shown in Figure 4.

Topic	Eval 1	Eval 2	Eval 3	Avg
Elian Gonzalez	88.2%	70.1%	75.0%	77.6%
Galileo Probe	82.6%	87.0%	91.3%	87.0%
Viruses	68.4%	73.7%	73.7%	71.9%
Vladimir Putin	90.3%	82.8%	94.7%	89.9%
West Bank	68.2%	77.3%	70.0%	72.3%
Average Precision:				79.1%

Table 1: A table showing the precision scores provided by each of the three evaluators on the five ASKNet networks.

In order to provide a more holistic test of ASKNet's network creation, large scale networks were produced from processing Wikipedia documents and the British National Corpus. These networks were then used as the basis for a semantic relatedness system, which determined the relatedness of two words by comparing the relative amount of activation received by each when the other was fired. This simple task, performed on networks created by ASKNet from general text, without any level of fine-tuning to the task specific goals was able to produce results comparable with the best performing purpose built systems[6].

2.4 BioMedical

ASKNet, and the tools which it uses have been developed for, and trained on newspaper text. It is true that these tools can not simply be applied directly to biomedical text without incurring significant loss to both their precision and recall. However, recent work has been done to begin porting the underlying tools to work in the biomedical domain. In particular, the C&C parser[3], which is the primary dependency parser used by the system, has recently been retrained to work with biomedical data, and was able to parse text from the BioInfer corpus with a precision of 81.4% [9].

Once the underlying tools have been adjusted to the new domain, the network formalism of ASKNet is robust enough to deal with most biomedical information without the need for significant adaptation.

3. NOVEL RELATION DISCOVERY

We have now seen that ASKNet is able to create large scale semantic networks, of reliably high quality, in short periods of time, and that the tools and structures used are capable of being ported to the biomedical domain. By combining the information integration of ASKNet with traditional NLP techniques, we can, for the first time, produce a system which not only retrieves and filters stored information, but that is also capable of discovering truly novel information from textual documents.

In the first stage of this system, we use ASKNet's spreading activation algorithms to produce a relational ranking; a list of entities, ranked by their relatedness to a target entity. As an example, we could produce a list of organizations ranked by their relatedness to a particular individual by simply adding activation to the node representing the individual, allowing the activation to spread, and then ranking the organization nodes by the amount of activation they received.

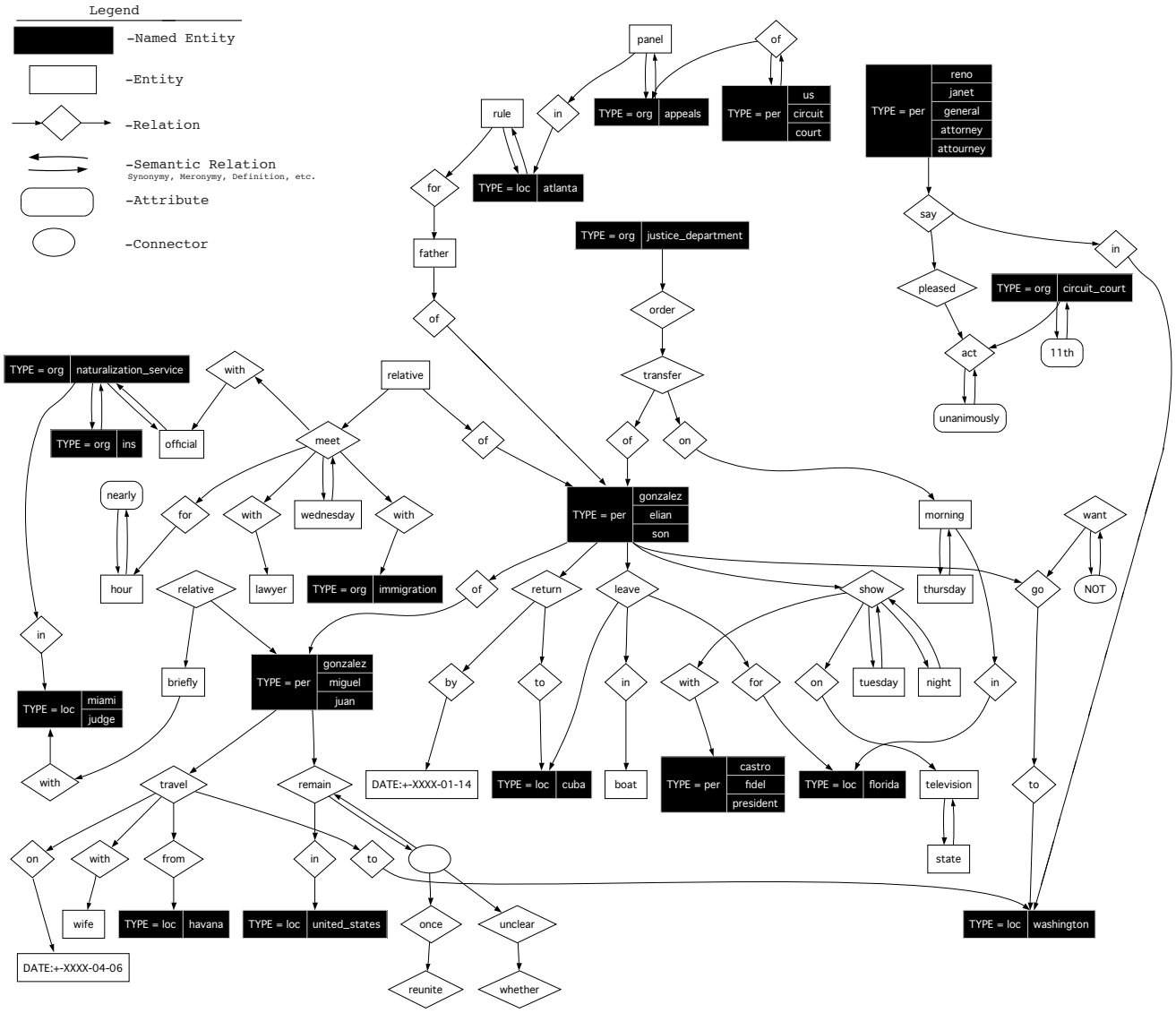

Figure 4: A sample of an ASKNet network created from multiple documents taken from the Document Understanding Conference. Nested structures were flattened for ease of human evaluation.

This relational ranking is in itself is a useful tool to biomedical researchers, as it could produce for example, a ranked list of chemicals most associated with a particular protein. However, most of the high ranking chemicals would be those that have the best known (i.e., most often repeated) relations with the given protein. So while this is useful for some tasks, its results are similar to those which could be found from a simple co-occurrence search. In our example ranking shown in Figure 5, it is entirely unhelpful to know that *Bill Clinton* is the person most related to *Bill Clinton*. It is also fairly obvious, and thus less beneficial to know that *Hillary Clinton* has a high index of relation with *Bill Clinton*. Both of these facts would have become obvious by performing a simpler co-occurrence search (as seen on the right hand column of Figure 5).

In order to discover novel relations, we must find high ranking relations, which are *not* accessible via a co-occurrence search. So in the second stage a simple co-occurrence search is preformed, and the results are removed from the relational ranking. All entities which have ever been mentioned in the same frame (which can be set to be an n-word window, sentence, paragraph or document), are removed from the ranking, producing a list of entities, ranked by their relatedness to the target entity, that have never been directly mentioned together with that target. In our biomedical example, this is equivalent to a list of chemicals, most associated with a particular protein that have never been mentioned together with that protein in the literature.

To understand the significance of this, we must revisit our example shown in Figure 2. The relational ranking system would be able to identify that there was a relationship between Chemical F and Disease B, but it would almost certainly be buried among thousands of other, more direct relationships that have been well covered by the literature. However, the novel fact discovery system could remove the extraneous relations, leaving only relations of exactly this type; those which require multi document analysis to discover, have never before been mentioned in the literature, and are thus the most likely to be true scientific discoveries.

In order to test this system, A network was produced from approximately 2 million sentences of text from the New York Times corpus. Bill Clinton was chosen as a target entity, as he was the individual most often mentioned in the corpus used. We preformed a relational ranking and a co-occurrence ranking on the *Bill Clinton* node, producing the two colums shown in Table 5. After excluding the co-occurrence rankings, we were left with several nodes, the highest ranking being *Richard Socarides*. Mr. Socarides was a white house adviser under Bill Clinton and held several high ranking positions in the Clinton administration. Clearly the two individuals were well acquainted, and worked together, but upon examination of the corpus used, it was confirmed that no single article directly linked the two.

This work is in its early stages, and currently only works on newspaper data, but the potential implications of a system which can automatically discover relationships between genes, proteins, chemicals and other entities in the biomedical domain are immediate and far reaching. This research has the potential to not only impact the way in which biomedical research is performed, but also to become the first completely autonomous system to make genuine novel scientific discoveries.

ASKNet	Correlational
Bill Clinton	Bill Clinton
Hillary Rodham Clinton	Kenneth Star
Moinca Lewinsky	Hillary Clinton
Al Gore	Monica Lewinsky
...	...
Richard Socarides	Linda Tripp
Mary Nell Lehnhard	Henry Hyde

Figure 5: Example output from the ASKNet relational ranking (*left*) and a simple correlational ranking (*right*)

4. REFERENCES

[1] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.

[2] J. Bos, S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier. Wide-coverage semantic representations from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1240–1246, Geneva, Switzerland, 2004.

[3] S. Clark and J. R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

[4] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975.

[5] B. Harrington. Managing uncertainty, importance and differing world-views in asknet semantic networks. In *Proceedings of the fourth IEEE International Conference on Semantic Computing*, Pittsburgh PA, USA, 2010.

[6] B. Harrington. A semantic network approach to measuring relatedness. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, 2010.

[7] B. Harrington and S. Clark. Asknet: Automated semantic knowledge network. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI'07)*, pages 889–894, Vancouver, Canada, 2007.

[8] B. Harrington and S. Clark. Asknet: Creating and evaluating large scale integrated semantic networks. *International Journal of Semantic Computing*, 2(3):343–364, 2009.

[9] L. Rimell and S. Clark. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, 42(5):852 – 865, 2009.

[10] P.-R. Wojtinnek, S. Pulman, and J. Völker. Automatically built semantic networks for determining semantic relations. *International Journal of Semantic Computing: Special Issue on Semantic Knowledge Representation*, 2011.