Managing Uncertainty, Importance and Differing World-Views in ASKNet Semantic Networks

Brian Harrington
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford, United Kingdom, OX1 3QD
Email: brian.harrington@comlab.ox.ac.uk
Website: www.brianharrington.net/asknet

***Abstract*—ASKNet is a tool for automatically creating large scale semantic networks from natural language texts. Its semantic network formalism is robust and extensible, but as with all formalisms of this type, it is limited in the types of information it can represent. In this paper, we extend the network formalism to better represent the strength of a relation as a combination of factors such as the relation's relevance, the reliability of the source, and our confidence that the relation is semantically correct. In addition, we extend the network to allow for post-creation adjustment of the world-view by turning off relations which were created by a particular source. This allows ASKNet to represent the information in its network as it would appear to a group or individual who either disregards, or does not have access to the information provided by a particular source.**

I. Introduction

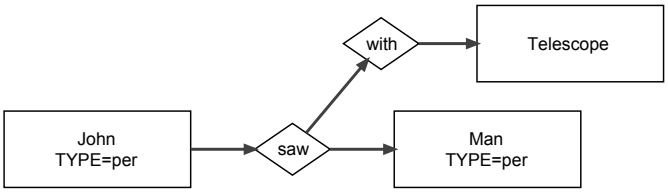
ASKNet is a system for automatically generating large scale semantic resources using information derived from natural language texts. Using a combination of existing natural language processing tools and a novel application of spreading activation, ASKNet builds semantic networks representing the information contained within a text, and then maps that information onto a larger network representing the sum of its world knowledge.

The semantic network formalism used by ASKNet is versatile and robust. A hierarchical nesting approach allows the network to represent a wide variety of information, without sacrificing simplicity and human ability to read and understand the networks. However, the spreading activation algorithms used by ASKNet rely on the relations having a "strength" parameter, which until now has been poorly defined and semantically ambiguous. It is clear that the strength of a relation bears some connection to that relation's level of certainty and importance, but precisely what this means in semantic terms has never been clearly defined. In order to rectify this problem, and also to allow users to more easily and intuitively manage the spreading activation algorithms, we have broken the strength of a relation into a number of different parameters, and added ways to manage each.

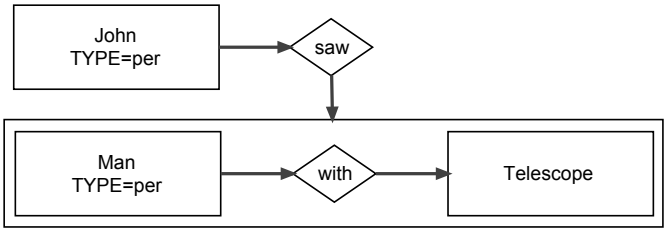
This work also introduces the concept of "sources" into the network. These sources can be used either as a means of adjusting the strength of the relations that they have authored, or to manage the world view of the network. Changing the world view of the network without having to re-build the network extends the usefulness of ASKNet, and opens some interesting avenues for further research.

II. The ASKNet Network

ASKNet uses a combination of natural language processing tools such as the C&C parser [1], and the semantic analysis tool Boxer [2] in order to produce discourse representation structures. These structures are then converted into semantic network fragments (see Figure 1), which are integrated into a single large network using a spreading activation based algorithm [3]. In our previous work we have shown that this method produces networks of a high quality [4], and that the resulting networks can be used to compute semantic similarity with results that are on par with the best purpose built systems [5]. An extract from an "real world" ASKNet network created from newspaper text is shown in Figure 2.



(a) John used the telescope to see the man



(b) John saw the man carrying the telescope

Fig. 1. ASKNet representations of the two parses of "John saw the man with the telescope"

The ASKNet network was designed for flexibility and ease of semantic understanding. The nested structure, as shown in Figure 3 allows complex nodes to be built out of combinations of existing nodes and relations.

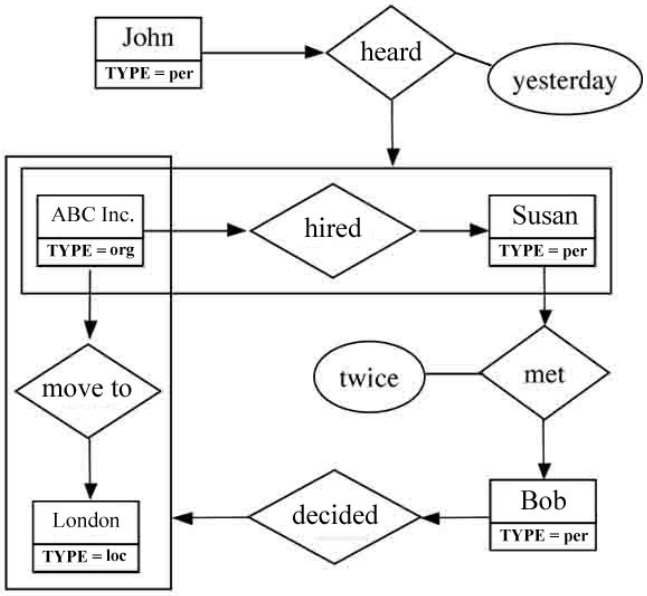


Fig. 3. A simplified Semantic Network created from the sentences "Yesterday John heard that ABC Inc. hired Susan. Bob decided that ABC Inc. will move to London. Susan met Bob twice."

Each relation in the network is named and directed, and multiple binary relations can be combined to form relations of higher arity. Each relation has an associated strength, which represents the certainty of the relation. These strength variables are important for the spreading activation algorithms, and improving them forms part of the basis for this work.

A. Spreading Activation

One of the key elements of ASKNet networks is their built-in spreading activation algorithms. These algorithms are essential in transforming the networks from a collection of small fragments based on individual sentences, into a large scale unified semantic knowledge resource [3], and can also be used in already built networks in order to perform a variety of tasks [4], [5].

In the spreading activation algorithms, nodes can receive activation from their neighbours via their relations. If a node receives an amount of activation that surpasses its firing threshold, the node will fire, sending its activation back through all of its relations to its neighbours. Spreading activation is based on research in psycho-linguistics [6], and has a long history in semantic networks [7], [8], [9], but was largely abandoned in the 1980s due to the intractability at the time of producing large scale networks on which it would be useful [10], [11]. Recent advancements in knowledge resources have brought about a resurgence in the use of spreading activation, mostly focusing on existing resources such as WordNet and Wikipedia [12], [13].

During the spreading activation algorithms, when one node fires, its activation is sent along all of its relations. The relative amount of activation sent along each relation is decided based upon the relations relative strengths. These strengths are values which represent the certainty or importance of a relation. A relation with a high strength value relative to others connected to the node will receive more activation, and will therefore be more influential to the outcome of the spreading activation algorithms.

As each node is fired, some of its activation is removed from the system. This process is known as signal attenuation. This is done to ensure that the system will always reach a stable state, and the amount of attenuation can be modified to control the rate at which the network stabilises, and the distance that activation can travel through the network.

III. Confidence, Reliability and Relevance

Semantically, the strength of a link can be interpreted in various ways. The higher the strength, the more influence the relation will have during the spreading activation algorithms. In previous work, this has been used to reflect a combination of factors, including confidence in the truth of the relation, the salience of a relation, or the reliability of the source from which the relation was derived [4]. This work separates and codifies these various factors into three parameters to allow for more intuitive and flexible manipulation of the network. In order to better understand the roles that each of these parameters play, we will first discuss how they contribute to the existing attenuation score, and then explain the changes that have been made to improve the way each is represented in the network.

The strength of a relation is now calculated from a combination of three factors: relevance, confidence and reliability, as described in Equation 1.

$$strength = \alpha\ relevance * confidence * reliability \quad (1)$$

Symbol Definitions

α	User defined parameter
$relevance$	Relevance factor as computed in Equation 2
$confidence$	Confidence factor as computed in Equation 3
$reliability$	Reliability factor as described in Section III-C

A. Relevance

Clearly some relations derived from text will be more relevant than others. A newspaper article discussing an important speech by a world leader, may make a passing remark to that leader's clothing to help set the tone for the reader, but we intuitively understand that the semantic relation between the leader and the topic of the speech is stronger than the one between the leader and the colour of the suit. Furthermore, there is a difference between permanent and temporary relations. For example, the sentences "London is located in England" and "London is very hot today" convey two different types of information. Initially, the ASKNet network took all relations to be permanent, and offered no way of allowing information to degrade. Additionally, while multiple occurrences of a relation would raise that relation's score, there was no way of adjusting the relevance aside from this crude metric.

In this work, we have updated ASKNet to allow relations to be set as either permanent, or degradable. This parameter is set at the time the relation is created based on the source material from which it was retrieved. This allows, for example, the user to set all of the relations created from an encyclopaedia as permanent, and relations created from a weather report as degradable.

Permanent relations act exactly as the relations had in previous versions of ASKNet, while degradable relations will have their degradation value slowly decrease, either based on signals sent by their source node (the source node will be discussed in greater detail below), or in reference to a global date variable. The date variable represents the date of the current article being processed, and degradable nodes check this date before each time they fire, and their signal attenuation variable is set accordingly.

If the user chooses to select a source article as degradable, which means that all relations it creates will be labelled as degradable, they then have the option of setting two parameters. The first variable is an expiry date, which can be set as a direct system date, or as an offset in number of days.[1]. The expiry date is the date by which the degradation value should reach 0. The user may also indicate whether the degradation should occur linearly or logarithmically. Therefore, if a degradable relation is created with a date of January 1st 2010, an expiration date of January 30th 2010, and a linear degradation flag; on system date January 15th 2010, the degradation value will be set to 0.5.

Permanent and degradable relations also reinforce differently. Permanent relations are reinforced as in previous iterations of ASKNet. This makes sense intuitively as hearing a fact from multiple sources should strengthen the relevance of that fact. How-

[1]All dates in this context refer to the date of the current article being processed. It is assumed that if degradable relations are being used, the "current" date is the latest date of any article processed so far. The date can also be overridden manually

ever, adding an additional copy of a degradable relation can either serve to "refresh" the relation, setting its degradation parameter back to the starting value, or in the case when the relation has degraded less than a user set parameter, the strength can be reinforced using the same method as is used for permanent relations.

$$relevance = \alpha\ deg + \beta\ (\frac{seen}{maxseen}) \qquad (2)$$

Symbol Definitions

α	User defined parameter
deg	Degradation coefficient *= 1 for permanent relations*
β	User defined parameter
$seen$	The number of times this relation has been reinforced
$maxseen$	The highest $seen$ value of any node in the network

B. Confidence

Intuitively, the strength of a relation should be dependent on how confident we are that it exists. In this context, we would like it to reflect both how confident we are that we correctly extracted the relation from the text (the *parsing confidence*), and how confident we are that the author of the text meant for that relation to hold (*the text confidence*). In earlier versions of ASKNet, the text confidence was not considered at all, and the parsing confidence was merely based on repetition. If a particular relation occurred multiple times in different documents, the strength of the relation was increased, which served to some extent as a form of parsing confidence.

Unfortunately, the C&C toolkit does not at the present time provide us with information on the probability of the parse that was generated (though we hope to work with the developers of the toolkit to provide this information in future). Therefore, the parse confidence, is generated through the use of a set of heuristics. In previous work we used human evaluators to analyse networks generated from text [4], from that information, we discovered common sources of error that could be attributed to the parser and semantic analyser. Heuristics were then created based on the frequency of errors caused by each source. For example, we found that the toolkit generated erroneous output on parenthetical statements in 40% of our test sentences. We therefore created a heuristic that reduced the parse confidence variable of parenthetical statements to 0.6. Admittedly, this is merely a rough stop-gap solution, until the parse confidence can be obtained directly from the toolkit, but it does provide the framework to easily implement an improved form of parse confidence when that functionality becomes available.

The text confidence value also uses a set of heuristics, but these heuristics are specifically left to the user to decide. A simple template has been created that allows users to supply a set of keywords that, if they appear in the text, will reduce the text confidence value by a set amount (also user defined) for all relations in the same sentence or paragraph. There is also a provision for the text confidence to be set based on the structure of the network. For example, a user can set a confidence value for all relations occurring within a "theme" parent node, which usually means that the relations are part of a quote within the text. Users can also set confidence levels for nodes linking to a negation node. However, users are specifically reminded that in many of the tasks for which ASKNet is used, something being explicitly stated as false can hold as much information as explicitly stating it to be true.

$$confidence = \alpha\ pars + \beta\ text \qquad (3)$$

Symbol Definitions

α	User defined parameter
$pars$	Parsing confidence coefficient
β	User defined parameter
$text$	Text confidence coefficient

C. Reliability

Not all documents are equally reliable. Likewise, not all relations should be trusted equally. In existing versions of ASKNet, it was possible to adjust the strength of relations on a per-document basis at creation time. However, as with the other factors, having a single value made it difficult to

understand the semantic implications of adjusting the number directly, and therefore difficult to determine meaningful values. It was also not possible to adjust the values for a particular source after the network was created. One can imagine a scenario where a particular source was determined to be less reliable than originally thought, and there would be no way of representing this in the network aside from rebuilding from scratch with adjusted values.

We have added two classes to ASKNet in order to deal with reliability. A *Document* class, which represents the document which provided the information to create a relation, and a *Source* class, which represents a source of information which generated documents. Each relation in a network is now linked to one or more documents, and each document is linked to one source. Since it is assumed that updating reliability measurements will be done less frequently than firing relations, the links are also maintained in the reverse direction to allow sources and documents to propagate news of a change of reliability measure, rather than forcing relations to look up their documents' reliability measures each time they fire. Each source object keeps a list of pointers to all documents it has authored, and each document object keeps a list of pointers to the relations it created.

At any time the user can change the reliability value of a particular source or document, and all affected relations will update their reliability value accordingly. By default, the reliability value of a relation is equal to that of its document which in turn is equal to that of its source. Changing a source's reliability changes all of its document's values by an equal amount, and changing a document's value changes the values of all of its associated nodes by an equal amount, except in the following instances:

- If a document's reliability value has been lowered by the user such that it is less than the value of its related source, changing the source's value will not update the document until such time as the source's value becomes lower than or equal to that of the document. This is done to allow a particular document to be set as uncharacteristically unreliable for a particular source, without violating the general rule that a document should not be more reliable than its source.
- Since a relation can be linked to multiple documents, its reliability should be equal to the highest reliability value of any of its documents. In order to maintain this invariant, when a document's reliability value is lowered, before decreasing the score of an associated relation, the relation's list of documents are checked, and its value set to the highest reliability value found.

IV. Managing World Views

Knowledge is not absolute, and therefore representation of knowledge should take this into account. Individuals can choose to believe or not to believe particular facts, and this will affect their world view. In the context of ASKNet, a world view is defined by the set of documents processed to build the global knowledge network. In pervious versions of ASKNet, the world view was set at the time of the network creation, and could not be altered. In this section, we will discuss alterations that allow a user to manipulate a network based on a particular world view, and discuss the implications of these changes.

Adding nodes for documents and sources not only improves the reliability calculations as described in the previous sections, it can also be used to manage different world views. In particular, it is now possible to adjust the reliability of sources according to the perceptions of a particular group or individual. This means that it is possible for a user to create a single global knowledge network, but then to work with that network in a form which correspond to the beliefs of a specific individual or group.

As an example, suppose we divide our sources into three categories; conservative, liberal and neutral. After building the network from information authored by all three groups, we could then choose to view the information in the network as it would appear to a person who only read or believed the liberal sources, by simply "turning off" the conservative and neutral sources, which would in turn send a message to their associated relations telling them not to fire. In the example shown in Figure 4, it would be possible to turn off either the liberal or the conservative source, removing either the grey or black nodes respectively. Note that if we turn

off the conservative source (as shown in Figure 5), we also turn off the relation connecting "Deepwater Horizon" to "Gulf of Mexico", and be left with an orphaned node, which would not be able to impact the rest of the network in any way. However the "in" node in "Oil spill in Gulf of Mexico" will remain active even though the conservative source has been turned off, as nodes remain active as long as any one of their documents has a source which has not been turned off.

Managing world views can have a variety of applications, such as "time shifting": removing all documents published after a particular date to view the information as it would have appeared on the given date, and "view based relatedness": using a similar technique to the one described in [5] to compare the semantic relatedness of nodes as they would appear to various people (e.g., for two individuals, if you know what sources they read, you could predict which one be more likely to relate "Barack Obama", and "Oil spill").

V. Conclusion

In this paper, we have improved the ASKNet network by replacing the simple strength parameter in relations with a series of parameters, which more accurately represent the level of certainty and importance the relation has within the network. This provides additional levels of flexibility for the users of the system, and divides the areas for further work into more manageable subsections. ASKNet is still very much a work in progress, and there are most likely other parameters which could be included to give an even more in-depth understanding of the strength of relations, but this work has provided a base for further development and consideration of what other factors may affect a relation's certainty or importance.

The ability to adjust the reliability of a source without rebuilding the entire network began as merely a way of improving our strength calculation. However, as a result of the changes made, it is now possible to turn portions of the network on and off, in order to represent information as it appears to individuals who either have not acquired, or do not believe information from a particular source. This seemingly simple modification has opened up several new areas for exploration, and could provide the basis for a good deal of future research.

The ASKNet network formalism has always been maintained with flexibility and extensibility in mind. In this work, we have shown how improving one aspect of the network can lead to a modification that allows ASKNet to represent a whole new type of information.

References

[1] S. Clark and J. R. Curran, "Wide-coverage efficient statistical parsing with CCG and log-linear models," *Computational Linguistics*, vol. 33, no. 4, pp. 493–552, 2007.

[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)*, Geneva, Switzerland, 2004, pp. 1240–1246.

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

[4] ——, "Asknet: Creating and evaluating large scale integrated semantic networks," *International Journal of Semantic Computing*, vol. 2, no. 3, pp. 343–364, 2009.

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

[6] D. Meyer and R. Schvaneveldt, "Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations," *Journal of Experimental Psychology*, vol. 90(2), pp. 227–234, 1971.

[7] M. R. Quillian, "The teachable language comprehender: A simulation program and theory of language," *Communications of the ACM*, vol. 12, no. 8, pp. 459 – 476, 1969.

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

[9] G. Hirst, *Semantic Interpretation and the Resolution of Ambiguity*, ser. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press, 1987.

[10] F. Crestani, "Application of spreading activation techniques in information retrieval," *Artificial Intelligence Review*, vol. 11, no. 6, pp. 453 – 482, Dec 1997.

[11] S. Preece, "A spreading activation model for information retrieval," Ph.D. dissertation, University of Illinois, Urbana, IL, 1981.

[12] H. Wang, X. Jiang, L.-T. Chia, and A.-H. Tan, "Ontology enhanced web image retrieval: aided by Wikipedia & spreading activation theory," in *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*. New York, NY, USA: ACM, 2008, pp. 195–201.

[13] V. Nastase, "Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, Honolulu, October 2008, pp. 763–772.

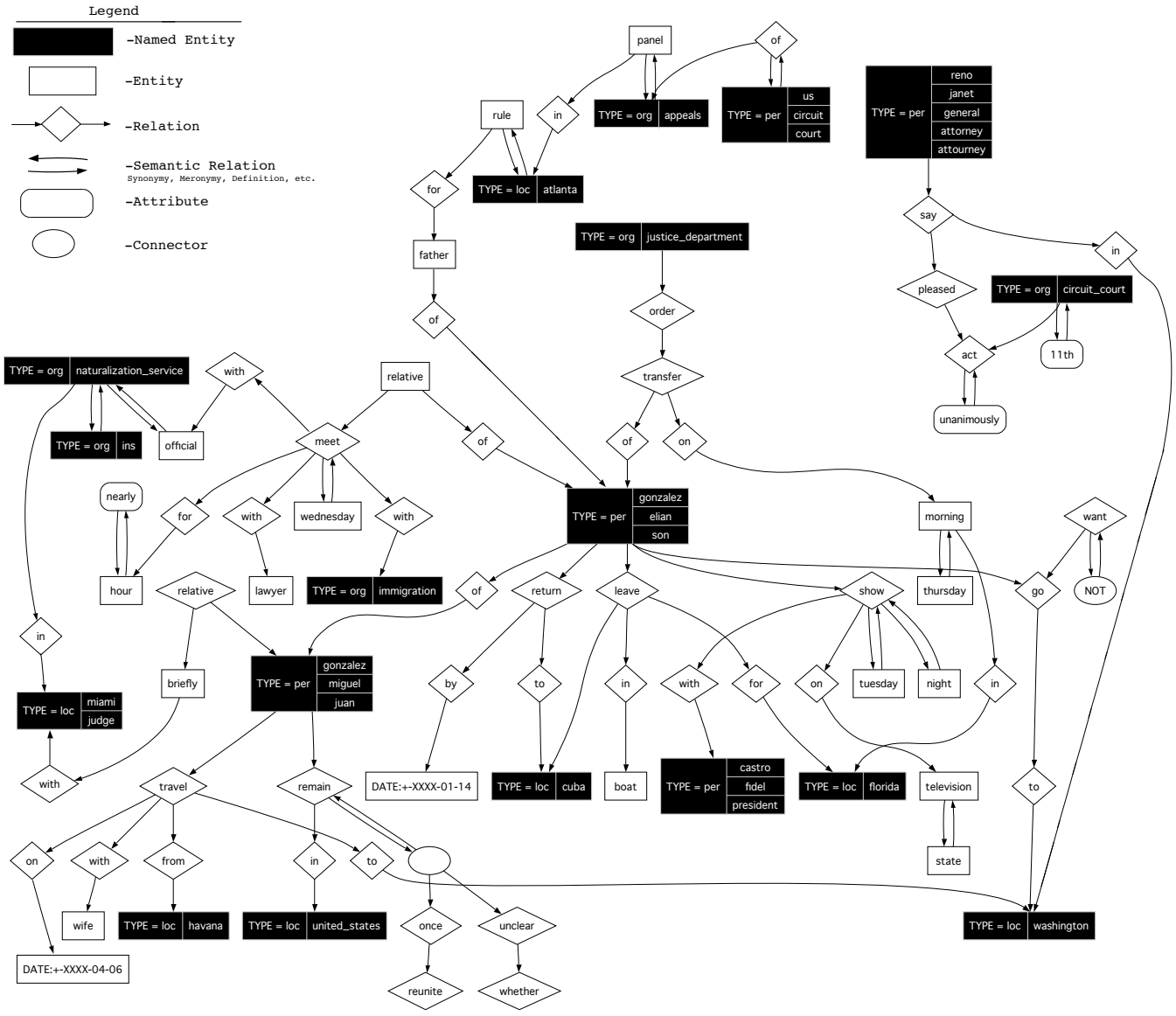


Fig. 2. An extract from a network created using newspaper articles regarding the 2000 Elian Gonzalez custody trial.

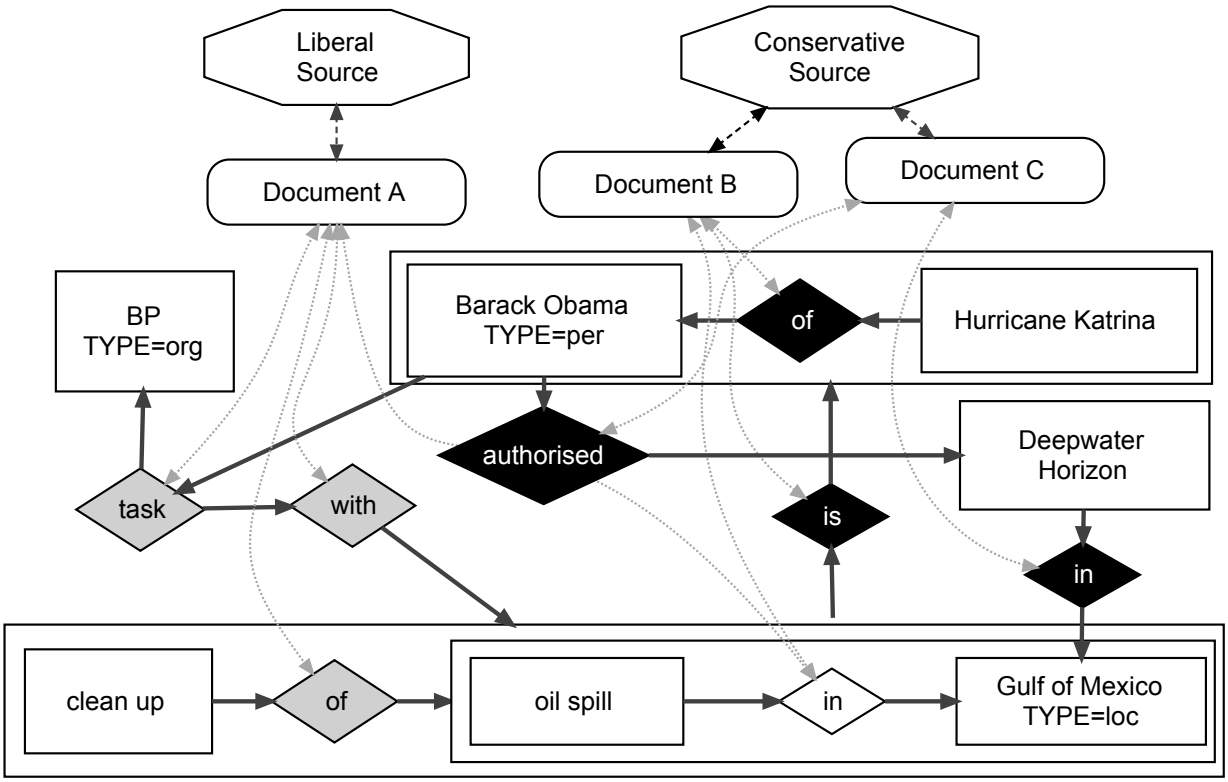


Fig. 4. An example ASKNet network showing source and document nodes.

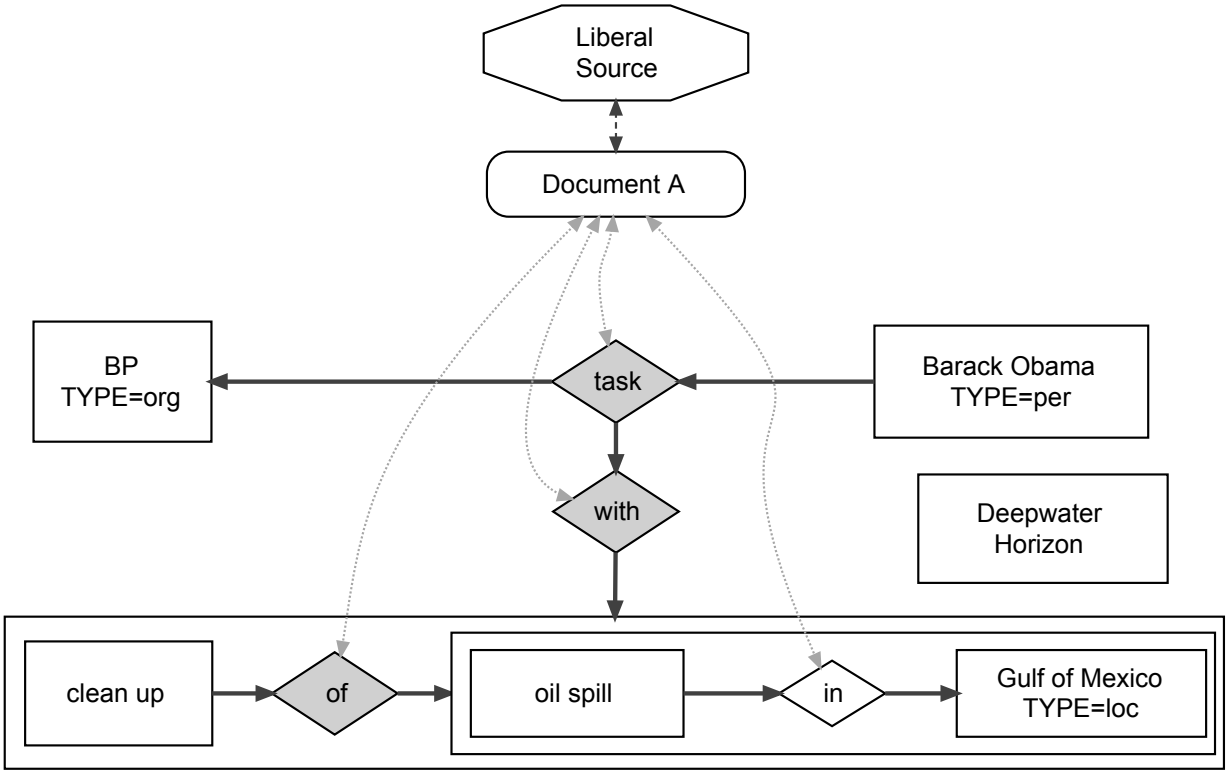


Fig. 5. The network in Figure 4 as it appears after switching off the Conservative source