Enhancement of unstructured and structured information

Trude Gentenaar, Job Tiel Groenestege*, Ronald Poell***

*Cognitieve Kunstmatige Intelligentie, Universiteit van Utrecht, Faculteit der Wijsbegeerte
**Netherlands Organisation for Applied Scientific Research (TNO)

This paper is submitted on 2005-04-04 for the CLIN 2004 Proceedings

Abstract

Information extraction from plain text is realized with the help of existing tools and specific software to create semantic network like structures — bridging the gap between an expression and the concept it stands for — which can than be added to an existing network. During this merging operation it is important to correctly identify matching concepts. The matching algorithm is composed of several complementary techniques (source model analysis, name matching and network context). We are aiming to automate these processes as much as possible.

1 Introduction

One of the major difficulties in information extraction is the interpretation of the provided content (and creating its semantics) and thus associating a meaning. Several frameworks are actually evolving in this domain e.g. RDF (Lassila and Swick) and Topic Maps (Pepper). The semantic network we created at TNO is another representation of real world knowledge that can be used in this area as a content reference network and will be explained in Section 2. Section 3 describes the information extraction process. Section 4 explains the process that integrates extracted information into the reference framework. The implementations of these processes are briefly described in Section 5. Section 7 concludes this article and Section 6, further improvements, looks ahead at the next steps in this field.

2 The Semantic Network of TNO

The domain independent Semantic Network developed at TNO is a network of concepts connected by meaningful relationships. It contains information from various sources representing some overlap. The original data models have been preserved and there has been no identification of identical content resulting in "double" nodes.

The network consists of nodes, relationships between nodes (statements) and attributes (properties) of nodes, statements and of attributes themselves (Poell).

The basic ideas about what the contents should look like go back to the late eighties and early nineties (Poell). Traditional databases models are conceived for a particular purpose, but in the Semantic Network content is added *whenever possible* and *as it is available*, without any presupposition on how and for what purpose it will be used. This results in a multi-model multi-author and domain independent

network in which the same kind of information can be available in different forms (and granularity). The mapping between the available models (ontology mapping) is not the responsibility of the creator of the content (early model binding) but of the applications that use the content for a particular purpose (late model binding). The rationale for this is based on our opinion that it is only the application that can define which view it should have on the available information. Some views might impose cardinality constraints[1], others might need high level abstractions.

The *easy way*, from our point of view the *wrong way*, is to put these application constraints in the information model. This is good practise in a controlled or closed application environment, however it does not hold for an open real world environment where things are not as nice and beautiful as normalized data would suggest (Poell).

3 Information extraction

3.1 Problem outline

This part of our research can be formulated as: *How can we get from text to a representation in the form of a semantic network?* To make this leap we need *(i)* the linguistic meaning of word phrases, *(ii)* the linguistic relationships between phrases, *(iii)* an interpretation of what each phrase is supposed to represent (meaning). Finally when placing this extracted information in a broader context it will be necessary to compare it to existing information (in this case to the information in the semantic network).

Although we aim ultimately to be able to correctly analyze the major part of a text, we do not have the pretention to be able to do so at this stage. With this work we hope to identify some of the next steps that will need to be taken. Figure 1 shows the kind of information that we would like to extract from the text given in Table 1. It is important to recognize that there is in general not one solution for the representation of the information contained in a text: *(i)* there might be different views of which groups of words represent the concept itself or *modifiers / attributes* of a concept *(ii)* sometimes a piece of information can been seen as *text* and will become an *attribute value* without representing a concept or represent a *concept* and participate in a relationship[2]. Fortunately in most cases there is one more or less *natural* way of representing the available information, but there is no *best* way as this refers to a specific exploitation context and the information extraction should be exploitation context free.

The information extraction process itself consist of three steps. First, semantic and syntactic annotation. Second, construction of a proto-network. Finally, the transformation in a semantic network. This network is then

[1] A specific person has only one birthdate but different sources can have different records of it. The information architecture (semantic network) allows the registration of all of them (of which only one is supposed to be true). Applications that have a view on this kind of data with a cardinality of 0 or 1 (no or one birthdate) have to decide which of them (the most likely) to present to the user. Intelligent services can assist the applications in this task.

[2] See the example of ditransitive verbs on page 5

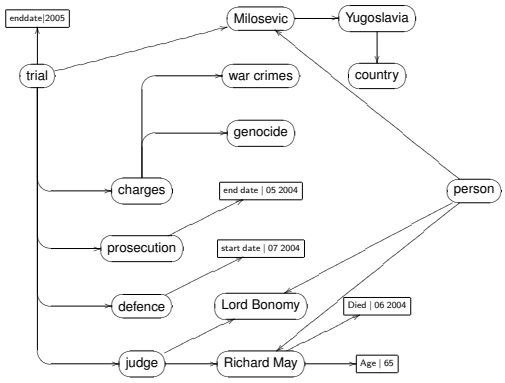


Figure 1: Example network from the news article in Table 1

Milosevic trial judge dies at 65 ***The judge who headed the trial of Slobodan Milosevic until earlier this year has died at the age of 65, the war crimes tribunal has announced.***
Richard May stood down from the UN tribunal at The Hague because of his ill health. He oversaw the lengthy first phase of the former Yugoslav president's trial on charges of war crimes and genocide. Judge May faced repeated defiance from Mr Milosevic, who refused to recognise the court's authority to try him. The judge sometimes switched Mr Milosevic's microphone off to silence the defendant.
Firm but fair In one of their most memorable early exchanges, Judge May asked Mr Milosevic whether he wished the full 32-page indictment, charging him with crimes committed in Kosovo in 1999, to be read out. Mr Milosevic told him: "That's your problem." In another hearing, Judge May told him: "Your views about the tribunal are now completely irrelevant, as far as these proceedings are concerned." The British former prosecutor was described as having an unflappable demeanour, with a reputation for being firm but fair. His decision to stand down from the court he had presided over since 1997 was announced in February and took effect in May after the prosecution wrapped up its case. He was replaced by another British judge, Lord Bonomy. Mr Milosevic is now preparing to start his defence later in July. The trial is expected to finish in 2005.

Table 1: Example of a news article

merged with the existing network with the help of the matching algorithms (see section 4). We utilize two tools to execute the steps. First, GATE as the framework (Cunningham, Maynard, Bontcheva, Tablan, Ursu, Dimitrov, Dowman, Aswani and Roberts) with ANNIE for the semantic annotation. Second, the Link Grammar Parser (Temperley, Sleator and Lafferty) to add syntactical structure.

3.2 Corpora

The used test corpus consists of 230 *English* articles from the BBC (288) and the New York Times (42). There are several reasons for taking news documents as the information source. As opposed to literary texts the assumption is that news is more directed towards factual information; this could help the extraction process. Moreover, the lexicon used by news reporters will probably be smaller than in literary text; this should be easier to annotate. Although scientific publications meet

these two observations, a big advantage of news texts is that they are grammatically a lot better than a scientific publications.

One more important reason is related to data that already resides in the Semantic Network. The network already contains a lot of geographical information. When using the *location* named entity it it should be easier for the concept matcher (section 4) to match it to existing structures.

Named Entities

Named Entity	total	mean
Date	1196	4.7
Location	2834	11.0
Money	166	0.65
Organization	4132	16.1
Person	2730	10.7

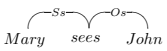
Table 2: Named Entity data of the BBC-news corpus.

A quick look at the annotations of the BBC corpus supports this idea, (Table 2). The mean locations per document, 11.0, seems even quite big, although there is one document with no less than 149 locations. A closer look (News) shows that this is indeed correct. The even higher mean of organizations per document can be explained by the fact that most documents contain the word 'BBC' one or more times.

The documents from the used corpora needed some preprocessing to get pure texts that does, up till now, require human intervention which is not conform the initial requirements.

3.3 Syntactical annotation

The Link Grammar Parser (LGP) parses sentences by creating links between every two words in a sentence. The transformation of a simple *Subject-Object-Verb* (SVO) sentence is straightforward; take the following linkage of the sentence *"Mary sees John"*[3]

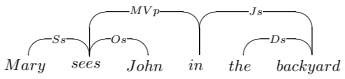


This can be converted into a statement:

The result is simple; two nodes and a predicate together form a statement. A slightly more complex can be obtained by extending the previous sentence with a modifier on the verb *sees*: *"Mary sees John in the backyard"*. The resulting

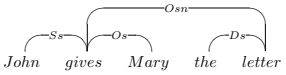
[3] Ss links a subject and verb, and Os the verb with its object. The small s says that it is singular.

linkage now becomes[4]:



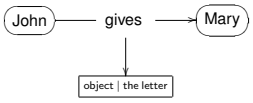
A more complex example contains a ditransitive verb such as *to give*. In a sentence like *"John gave Mary the letter"* a simple statement will not be possible as there are two objects: a direct one (*the letter*) as well as an indirect one (*Mary*).

The linkage of this sentence is the following[5]:

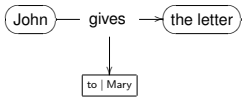


Notice that *to give* does not necessarily have two objects. It can also be used in a normal transitive way (*John gave the book*). There are five different ways to

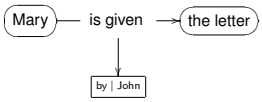
Ditransitive



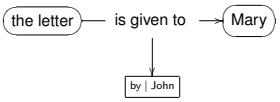
"John gives Mary the letter."



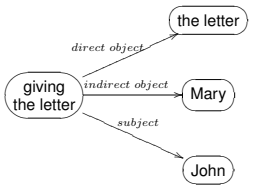
"John gives the letter to Mary."



"Mary is given the letter by John."



"The letter is given to Mary by John."



topic-map representation

Figure 2: Different ways to represent the meaning of a sentence with a ditransitive verb.

[4]The `MVp` is the link that connects verbs and adjectives to modifying phrases. `Js` connects prepositions to their object.

[5]The `O*n` link of the linkage is always the second object. One of the reasons that it is a different is to prevent linking two pronouns to the verb as an object "*John gave her it" or "*John gave Mary it".

represent a ditransitive verb in the network. These are shown in Figure 3. The first form is the representation that needs the least transformation from its syntactical form to the network representation but all five representations express the same action of giving. However, reading them back from their network representation they seem to have a difference in their *Natural Language* (NL) counterpart. The topic map representation is the only one that has not one natural way of *reading* the provided information.

3.4 Construction of a semantic network

At this point, the extraction process can be seen as a way find to put relations between clusters of annotations. The relations are primarily determined by the grammatical structure and the clusters are resolved by using the syntax as well as the semantic annotations of GATE.

The actual extraction process is split up in two main stages. First, we create the proto-network that will only contain partial statements and attributes. Second, a Semantic Network is created from the portions of the proto-network that are found to be complete statements and attributes.

Extracting or clustering annotations into the proto-network is done through sets of patterns executed in a predefined order. Every pattern set uses specific types of annotations. Each pattern is comprised of a rule that tests wether the pattern applies and a procedure that modifies the proto-network by adding or changing elements. The application of the patterns in a certain set is done by applying each pattern in that set on every annotation that is applicable to the set.

The creation of the Semantic Network from the proto-network can, again, be divided into several stages. First, a Semantic Network structure is created from the proto-network. Only predicates and attribute-types that form complete statements and attributes will be used. Second, the newly created nodes are named using the annotation cluster and the text they annotate. Finally, this new structure is uploaded to the Semantic Network server.

4 Matching process

4.1 Problem outline

The Semantic Network as created at TNO contains the contents of several (partially) imported sources [6]. Individual authors can edit parts according to their own view on the world.

Separate representations can cover the same domains, resulting in nodes representing the same concepts. The semantic networks resulting from the information extraction in section 3 can be seen as another source to be imported in the net-

[6] The most important are Notion System(Poell), WordNet(Miller, Fellbaum, Tengi, Wolff, Wakefield, Langone and Haskell), Nima, OpenCyc(Ope) and the TNO Phsics and Electronics Laboratory Intranet. The network consist actually of 1,2 million nodes and 4 million relationships.

work[7]. Ideally all these types of doubles should be identified and, when appropriate, merged.

The goal of this research is to create a procedure that can identify correspondences between knowledge representations and make its conclusions in a transparent fashion. The final result of this procedure need only be the *correct identification of double concepts*. The identical or matching concepts are then known to applications that will make use of this knowledge.

4.2 Different representations and automatic analysis thereof

The original sources from which the Semantic Network has been build are mostly either relational databases or semantic networks, notably absent are any of the traditional ontologies (SUMO (Niles and Pease), etc). The reason for this is that ontologies usually only provide an explicit specification of a conceptualization (Gruber), not the contents of said conceptualization.

Databases are usually very domain dependent with no linkage to other domains. Semantic networks on the other hand can be domain independent, but both can offer a large number of specific instances. The ways in which semantic networks can differ is described in (Klein). There can be big differences in granularity. For example when regarding geographical entities, WordNet states the larger cities of for example France where Notion System (Poell) lists all the provinces, departments and all of the villages as well as the larger cities, see figure 3. Most

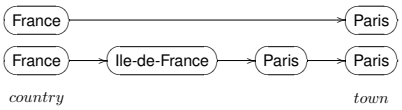


Figure 3: Difference between two data structures

of the representations differ more or less from each other in scope and granularity but the output from the information extraction procedure has a structure different from anything seen so far. There are fewer statements than the previous representations and a complete lack of a taxonomic hierarchy. Because of these marked differences these representations can not be dealt with in the same way as the other representations.

The differences in knowledge representations provided here is far from complete but it covers the ground that is relevant to this research.

[7]The matching process described here can be applied before actually adding this *temporary* network to the existing one or as a post import process. The form of these networks are slightly different and not all of the applied methods are as effective.

4.3 Analysis of sources

The nodes imported from various sources always contain a so called *identifying attribute*, indicating that at least a part of this node comes from a specific and recognized source. Generic templates are generated using these ID attributes.(see Figure 4). Each generic template contains three clusters: *(i)* used attributes, *(ii)*

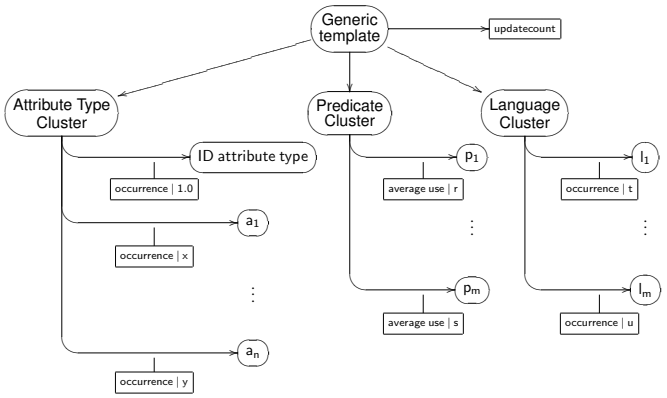


Figure 4: A generic template

used predicates, and *(iii)* the languages used in the source. The average number of times a predicate is used in a node is logged in the predicate cluster. The attribute and language clusters represent the occurrence of these elements encountered in each source in the range: >0 to 1 (1= they always occur).

These templates are generated from a representative random set of nodes for each source. A statistical analysis is carried out on the presence of statements (their predicates) and attributes of the nodes and the template is updated accordingly [8]. It is not necessary that all of the predicates and attribute types that are used within a certain source are seen. We are interested in learning about the most important, most extensively used predicates in a source.

The difference in the number of predicates between Notion System and WordNet (Table 3), is interesting. As these representations are both domain independent, the predicates in Notion System can be said to carry *more specific information* than the predicates used in WordNet. In Notion System predicates shall be called *descriptive* whereas the those from WordNet are said to be more *general*. These are relative terms, it is not possible to make absolute statements about the descriptiveness of predicates. The restricted number of predicates used in the other sources (NIMA, FEL) is related to the restricted domain these sources cover.

[8]The use of attributes on statements and on other attributes is not registered by the generic templates at this time. It was deemed that this information would not benefit the purpose of the generic templates enough to account for the cost of gathering this information.

	WordNet	Notion System	NIMA	FEL Kennis ID	FEL project nr
update count	2000	5500	2500	75	150
predicates	14	90	2	7	4
attributes	4	65	7	3	5
languages	1	23	5	1	1
# nodes	200000	210000	700000	1800	6300

Table 3: Results of the templates

Together with the creation of the generic templates, specific templates are created, these can be used to distinguish different *kinds* of nodes. There are two different kinds of specific templates based on the usage of *descriptive predicates* or *general predicates*. A specific template has exactly the same form as a generic template as in Figure 4 but with less elements in each cluster. A new specific template is created each time a node scrutinized contains a certain number[9] of new predicates and attributes not present in the existing specific templates. This method cannot be applied to sources that use only general predicates (as there are too few of them). To be able to form meaningful specific templates the chain of nodes formed by following relationships with one predicate is used[10]. In WordNet e.g. the is a(n) predicate was used, leading to a chain for *The Netherlands* of: European country, country, administrative district, district, region, location, object, entity. Because using this entire chain would result in to many, to specific templates, only the top three nodes (location, object, entity in this case) are used. Before these templates can be used they need to be cross-linked (specific templates from one source are linked to specific templates from other sources that represent the same kinds of nodes). This cross-linking is realized by providing matching nodes in different sources.

4.4 Mapping procedure

The mapping procedure starts with one pair of nodes the *start pair*[11]. From these nodes the *Pairwise Network Crawler* or PNC will gather possible other matching pairs and finally *Judging modules* will evaluate if two nodes match or not.

Pairwise Network Crawler The underlying assumption of the PNC is that when two nodes are identified as representing the same concept, the likelihood of finding more matching nodes in the neighborhood of those two nodes is increased. When

[9] The threshold for this number has been determined experimentally.

[10] The selection of this predicate is done for now by hand but future work will enable this automatically.

[11] Start pairs can be chosen by hand and are known to be correct matchings. The network might also be compared exhaustively but this is a resource consuming task. The last possibility we will mention here is looking in the network for the matchings of one specific node. In this case a name search will provide an initial set of start pairs.

two nodes match, their neighbors might just match as well. The PNC will analyze the network related to each of the nodes in a start pair and generate pairs of candidate matching nodes as a cartesian product of the retained nodes.

In order to avoid an explosive search when following all the relationships, two kinds of restrictions are used: *(i)* predicate classification and *(ii)* limits in network depth used. Further restrictions in establishing the list of candidate pairs are applied trough name-based filtering.

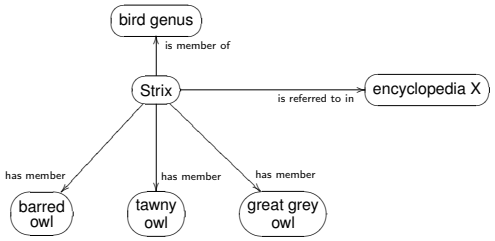


Figure 5: Three different types of predicates: is member of is of the generalizing type, has member is of the specifying type and is referred to in is classified as an other predicate.

The predicates are divided into three different categories. Figure 5 shows a node where these three different types are illustrated. The three types of predicates are the *generalizing predicates* (hypernyms and holonyms), the *specifying predicates* (hyponyms and meronyms), and the *other predicates*. In most cases a generalizing predicate will relate a node to only one other node and the network depth allowed is not restricted. Specifying predicates occur often more than once for a particular node and the PNC is limited to a depth of 2 or 3. For relationships with the third kind of predicates (other predicates) the allowed depth is only 1 (direct related nodes) as there are no reasons to believe that deeper related nodes might have a logically matching node in both branches.

The last reduction of the list of candidate matching nodes is realized through a name-based filtering process. The retained candidates will have corresponding names exceeding a defined threshold. Three string comparison methods were tested: *(i)* the standard Java equals function, *(ii)* the Levenshtein or edit distance (Gilleland) and finally *(iii)* the algorithm that was found to perform the best in this situation can be found in the following article (White) [12].

Table 4 shows start pairs consisting of one WordNet node and one Notion System node. In the second and third column the number of nodes reachable from each of the start nodes is shown, in the next column the number of pairs that were

[12]Both common substrings and common ordering of those substrings are rewarded. This algorithm returns a value between 1 and 0 that indicate the similarity between strings. The PNC compares this value to its threshold to judge if two strings are similar enough to retain the associated pair of nodes. See for a more detailed explanation of the algorithms and rational (Gentenaar and Tiel Groenestege)

	WordNet	Notion System	pairs
Africa	151	389	77
America	613	32295	1583
Asia	243	379	52
Australia	31	54	8
Europe	424	43625	215
Animals			974
Plants	9018	6219	364

Table 4: Results of the Pairwise Network Crawler, the numbers in the reachable nodes columns are minimums

found are shown. From this table we can see the density of matchings, for example, the start pair Australia leads to 31 nodes in WordNet and to 54 nodes in Notion System and 8 of those match. This seems like a reasonable result. For America we see that they do not add up, there are only 613 nodes reachable from the WordNet start node in this pair but there are 1583 matchings. A large number of these matchings are not correct and it is up to the judging modules to find out what matchings are wrong. Another thing that can be seen from Table 4 is the difference in granularity. Notion System has a far greater number of locations per continent than WordNet who does not go into the same level of detail as Notion System when it comes to the smaller towns within countries.

Example In Figure 6 the start pair (two *Strix* nodes) is extended to form a total of four sets, in this example there are only generalizing and specifying predicates. The nodes connected to the specifying predicates are collected in the sets s_1 and s_2. The sets g_1 and g_2 comprise of the nodes connected to the generalizing predicates and are extended to include their complete paths.

Once the sets have been determined all the nodes in set g_1 are paired of with the nodes in set g_2 and the same goes for the sets s_1 and s_2. In this results in 169 pairs. Finally these 169 pairs are compared on their names, 8 pairs of nodes have names that are similar enough to pass through the filter, they are shown in Figure 6.

Remarks Due to the restriction to limit the depth (network distance) along specifying predicates, it is not possible to locate nodes with the same proper names who's representations differ by more than this distance.

Another drawback in this procedure lies with the character string matching. When less exact matchings are allowed more potential pairs are created. When the density of matchings is found to be low, less strict string matching is allowed. The adjustment of the strictness of the string matching filter is done by hand for now, however it should be possible to automate this adjustment in the future.

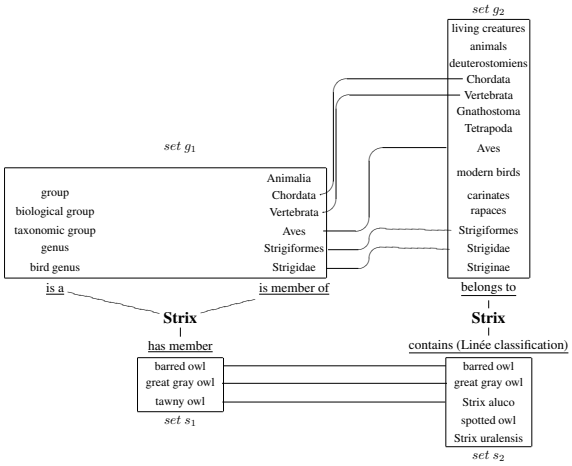


Figure 6: Sets of nodes and the pairs after filtering

Alternative approach The source networks resulting from the information extraction procedure are very different from the ontologies, databases, and semantic networks seen so far and a different approach is needed. The procedure as described sofar does not actually work on this type of source. However some of the techniques used here can be utilized to make a start at locating matchings for nodes from this source [13]. The nodes generated from a single article are always treated as a set of nodes belonging together. The names of each of the nodes in this set are used to search the entire Semantic Network. The result is then used as the set of potentially matching nodes to the node that supplied the search query. It is up to the judging procedure to pick the best option [14].

Judging modules Each judging module (Specific Template Matcher, Network Distance and Semantic Distance Evaluator and Context Index Evaluator) looks at a certain aspect of a potential pair and the result of the analysis is stored and used in the final decision making procedure (Judge).

Specific Template Matcher This module compares the node scrutinized to the available specific templates. For templates based upon descriptive predicates the same formula is used as for the decision to create a new specific template or

[13] In the case of a knowledge representation with a well formed network structure, one start pair can be enough to initiate a run through the network resulting in a few hundred potential pairs. The nodes of extracted information are lacking this mature network structure and so a network crawl is not fruitful.

[14] In this case there is no possibility to generate more potential pairs from this start pair so we move straight on to the decision making process.

not. For templates based upon general predicates the result is either 1 or 0 (having the same classification levels or not).

Network Distance and Semantic Distance Calculator When all the pairs that do not fit into a template pair have been filtered out, there is still the possibility of a single node having more than one potential match. This module (together with the Context Index Evaluator) was designed to differentiate between these different matchings.

The *network distance* between two nodes in a semantic network is equal to the minimum of statements that need to be traversed to get from one node to the other.

The *semantic distance* between two nodes is equal to the minimum number of *not semantically neutral statements* that need to be traversed in order to get from one node to the other or 1 if only semantically neutral relationships occur. Semantically neutral statements are transitive and their predicates are either identical or semantically equivalent[15].

Figure 7: A small sample of a semantic network

In Figure 7 The network distance between Patricia and the Earth is 4, the semantic distance is 1. The chain of statements from San Francisco to the Earth use only semantically neutral predicates. This means that the statement "Patricia was born in" can be virtually linked to any of the nodes in the connected chain.

The Network Distance and Semantic Distance Calculator will determine the network distance and semantic distance for those groups of pairs that have one node in common.

Context Index Evaluator As we have seen the result of the information extraction procedure is a very immature network. Because of this the previous modules are not useful. This module uses the assumption that all the nodes that were extracted from a single document have an implicit relationship with each other ("originated from" a certain article) and this is enough assist us in formulating another measure of likeness.

For this module only the extracted nouns, adjectives or adverbes are considered. These nodes often result in more than one suggestion from the abundant WordNet source. As this source often contains multiple senses for a single word, this module needs to decide which sense was intended in the original article. To do this all the words that made it through the extraction process and are of the right

[15]The semantic distance needs to reflect the semantic relation between two nodes. A small semantic distance indicates a strong relationship between two concepts. See (Poell) modified

type. All the other words were excluded as they do not carry the type of information that is relevant at this point. For each of the WordNet nodes that are found as potential candidates all the words used to describe the sense are gathered as well as the names of all the nodes that it is connected to.

What follows is a rather crude but effective counting of all the words that occur in both sets. The WordNet node with the most hits is the most likely candidate for matching to the node from this particular article.

Judge The final conclusion is reached using the a few steps (Figure 8).

The most important criterion that the potential pairs need to meet is the fitting into a pair of templates. This then constitutes the first filter, any pairs that do not fit the templates are discarded at this point.

The set of potential pairs that remain after the first filter can contain double nodes (a node appears in more than one potential pair). The best pair is selected by looking at the results of the semantic- and network distance calculation. The pair with the lowest value for both measures is selected, the rest is discarded.

In the case of the extracted information networks, the context index is used to differentiate between double nodes. The matching node with the highest score is selected, the rest is discarded.

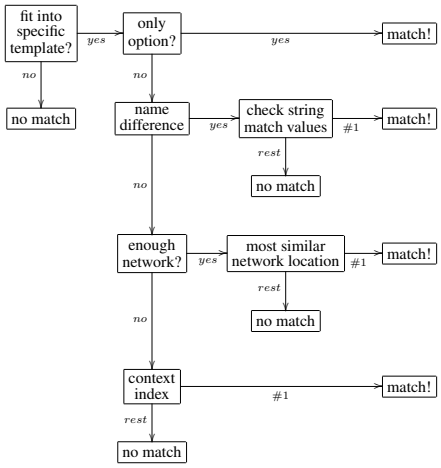


Figure 8: The decision making process

5 Implementation

The Link Grammar Parser (written in C) has been integrated as a processing resource in GATE (written in Java) and interfaced by the Java Native Interface (JNI).

For the matching procedure all the software is written in Java as this facilitates the interaction with the Semantic Network Engine itself written in Java[16].

6 Further improvements

The information extraction part of this research highly depends in several stages upon the existing Link Grammar Parser and GATE tools. Improvement of these tools, in particular other languages for the LGP and GATE's co-reference matcher, might immediately improve the obtained results.

Other or improved ways of generating the prototype-network from linkages would extend the kind of information that can be extracted.

For networks resulting from information extraction, network clustering techniques (See (Li)) can be applied to isolate tightly interconnected parts of the network that provide a more complete information context than the WordNet description attribute.

The realization of an *objective* quality measure of the extracted information is still a task to fulfil. This measurement should contain an evaluation of how much of the total amount of text is represented in the prototype-network and finally in the generated semantic network.

A *subjective* useful measurement would be an evaluation of how different the generated network is from a network based on the same information but generated by human beings. This measurement is necessarily subjective because different persons might or might not create different networks from the same information.

Last, but certainly not least, some of the empiric thresholds and predefinitions that are actually hand-made could be build up from scratch by a automatic learning process.

7 Conclusions

Although we are not yet able to extract from texts the information we would like to, parts of the extracted semantic networks are already valuable and there is still room for improvements. The matching of nodes in an mature semantic network fulfils almost completely its role. For immature networks, like the ones resulting from information extraction, new modules have to be realized to reach the same level of satisfaction.

The results of this research are extremely promising for a better automatic analysis and usage of textual documents in a Semantic Network and Semantic Web information environment.

8 Acknowledgements

This work has been conducted as a Phd thesis of the University of Utrecht in collaboration with the Netherlands Organization of Applied Scientific Research.

[16]For complete details about the implementation see (Gentenaar and Tiel Groenestege)

References

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N. and Roberts, I.(2001), Developing language processing components with gate, `http://gate.ac.uk/sale/tao/index.html`.

Gentenaar, T. and Tiel Groenestege, J.(2005), *Enhancing Information. Information Extraction and Concept Matching using the Semantic Network Engine*, PhD thesis, University of Utrecht, Netherlands organisation for Applied Scientific Research.

Gilleland, M.(n.d.), Levenshtein distance, in three flavors, `http://www.merriampark.com/ld.htm`.

Gruber, T. R.(1993), A translation approach to portable ontology specifications.

Klein, M.(2001), Combining and relating ontologies: an analysis of problems and solutions, *in* A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt and M. Uschold (eds), *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA.

Lassila, O. and Swick, R.(1998), Resource description framework (rdf) model and syntax specification.

Li, Q.(2003), *Graph-based Clustering Approaches for Semantic Networks*, PhD thesis, Technical University Delft, Netherlands organisation for Applied Scientific Research.

Miller, P. G. A., Fellbaum, D. C., Tengi, R., Wolff, S., Wakefield, P., Langone, H. and Haskell, B.(n.d.), Wordnet, `http://wordnet.princeton.edu/`.

News, B.(2004), Should foreign troops leave lebanon?, `http://news.bbc.co.uk/2/hi/middle_east/3624418.stm`.

Niles, I. and Pease, A.(2001), Towards a standard upper ontology, *in* C. Welty and B. Smith (eds), *2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.

Ope(n.d.), Opencyc, `http://www.opencyc.org/`.

Pepper, S.(2000), The tao of topic maps - finding the way in the age of infoglut.

Poell, R.(2001a), Notion system, `http://www.notionsystem.com`.

Poell, R.(2001b), The semantic network of ikm-i3, `http://www2.gca.org/knowledgetechnologies/2001/proceedings/Poell\%20Slides.ppt`.

Poell, R.(2002), Semantic network, part 1: Introduction and user's point of view, Draft Internal TNO report.

Poell, R.(2005), Information phylosophy for the semantic network, whitepaper.

Temperley, D., Sleator, D. and Lafferty, J.(n.d.), Link grammar parser, `http://www.link.cs.cmu.edu/link/`.

White, S.(2004), How to strike a match, `http://www.devarticles.com/c/a/Development-Cycles/How-to-Strike-a-Match`.