An Application of the Semantic Web Inspired by Human Learning and Natural Language Processing

R. Spiegel[1,2,3]

[1] Ludwig-Maximilians-Universität München, Institut für Medizinische Psychologie, Munich, Germany
[2] Technische Universität München, Medizinische Fakultät, Munich, Germany
[3] University of Cambridge, Wolfson College, Cambridge, United Kingdom

***Abstract*—The prototype in this paper presents a semantic web application that is inspired from psychology experiments on human learning and natural language processing. It aims at improving the proficiency of present search engines when dealing with specific queries (question-answering). The prototype makes use of the idea that the world wide web itself contains an enormous number of documents written in natural language (appearing in formats such as .html, .xml, .cfm, .pdf, .net, .asp etc.). It is consistently trained to improve its language proficiency by extracting knowledge from these documents and by storing redundant information in a database (i.e. information dealing with the same concept but expressed in different words).**

***Index Terms*—Semantic Web, Human Computer Interaction, Psychology-based application, Simulation.**

I. Introduction

The burgeoning role of the semantic web and its applications is mirrored in a recent *IEEE Computer* issue [1]-[2]. With the growing use of search engines throughout the ever expanding online community, fast and efficient search has become a major criterion. Today, appropriate information is mostly sought by typing keywords into Google, Yahoo, AskJeeves etc. These search engines are fast at providing a large number of pretty relevant search results. In many cases, however, users request an answer to a very specific question, e.g. when aiming to find out how many people used to live in Vienna in 1879, one might be tempted to ask "what was the population of Vienna in 1879?" Unfortunately, present search engines cannot deal with questions of this type. One reason why they struggle is because they do not understand the question in the first place. Taking inspiration from past question-answering systems [3] and combining the progress in this field with scientific evidence on human learning and language acquisition [4]-[9], this paper aims to provide a solution to this problem. It is important to keep in mind that such a system requires a number of features for this purpose. First, it needs to be able to understand the query. In order to understand the query, however, it not only needs language processing elements that are able to deal with grammatical rules and its exceptions, but also a large base of existing knowledge that it can refer to in order to understand the meaning of the query. In other words, the system needs to be able to deal with both syntax and semantics. This will assist it to first understand the meaning of a query and to provide an appropriate answer subsequently. Moreover, the system should be able to update its knowledge base continuously, as information changes quickly, e.g. when asking the system "what team has won the most recent Fifa World Cup", the correct reply would be "Italy" at present but would have been "Brazil" several months ago. By enabling the system to communicate in such a way, this process may ultimately lead to a more refined internet search in the future. So far, however, a system combining these features seems out of reach. Although Google and Yahoo have enjoyed growing popularity, their present limitations are illustrated by Lotfi Zadeh's research on how well they interpret questions [10]. Asking Google "what is the population of New York" generated results such as "News results for population of New York – View today's top stories…after the twin tower nightmare, New York is back on form, says…UN: World's population is aging rapidly – New, deadly threat of Aids virus…" As long as a system struggles to interpret queries in the way people would interpret them and as long as this system provides answers that fail to relate to the meaning of the question, there seems little hope for groundbreaking solutions to this problem. Consequently, the prototype presented in this paper focuses on teaching the system syntax and semantics of a given language. Now one might ask how to teach a system natural language in a way that it is able to communicate effectively with its users. Obviously, present search engines and other artificial systems lack a comprehension of common sense knowledge (though fruitful approaches are underway to overcome this problem, e.g. Doug Lenat and his Cycorp colleagues have teamed up to teach artificial systems common sense knowledge rendered in a formal language. Since 1984, they have been entering phrases representing common sense knowledge such as "water is wet", "every person has a mother" [11]). The idea put forward in this paper is very similar, but instead of teaching the system common sense knowledge and the formal rules of grammar by manually entering phrases, this paper aims at enabling the system to train itself by automatically extracting common sense knowledge from the World Wide Web. The advantage of automatic training lies in the system being able to make use of a larger training set. Moreover, automatic training is less time-consuming. When Doug Lenat started Cycorp back in 1984, the internet was mostly used for military and researchoriented purposes. Consequently, there was hardly enough common sense knowledge on the web that could have acted as a training set. Nowadays, however, the web itself is a huge (though sometimes unstructured) network representing both common sense and more specialist knowledge. As a result, my idea is to extract knowledge in order to let the system train itself. One caveat remains in order, though: Whilst Doug Lenat and his colleagues at Cycorp were able to explicitly teach the system knowledge rendered in a formal language, the information on the web is not rendered in a formal language yet (apart from the programming language of the website, e.g. .html, .xml, .asp, .cfm etc.). Thus, a system like the one proposed here would not only automate the process of extracting information, but also automate the process of developing a formal language that is the same for all extracted phrases. Ultimately, both systems (i.e. the one proposed by Doug Lenat / Cycorp and the one proposed in this paper) have the same aim, i.e. being able to actually *understand* language. This is a relatively challenging task. In order to solve this problem, a look at linguistics might help. Referring to Noam Chomsky's rationalist idea of language competence might be a starting point [12]-[13]. In order to understand a language, it is not enough to know about this language's vocabulary, its grammar rules and its exceptions, as the same words can be open to different interpretations. Consider the following examples [13]: "Mary expects to feed herself" / "I wonder who Mary expects to feed herself." In the first example, "herself" is clearly related to "Mary", whilst "herself" is related to "who" in the second example. Human beings are able to easily understand expressions of this kind. So far, however, artificial systems struggle with these problems. Moreover, this example can be extended to an almost infinite number of other examples, just by replacing the name "Mary" by another name or another substantive, such as "the tailor", "the cat", etc. Moreover, the verb could be replaced by another verb, "who" could be replaced by "that" and "herself" could be replaced by "himself" or "itself." This is not even where the story ends, as these words can appear in singular or plural form, not to mention the many other phrases where just one word changes both meaning and grammaticality. To illustrate this, I have chosen another example that is not related to the earlier example, but poses similar problems: e.g. "who do you think you are?" vs. "who do you think you are supporting?" As can be seen, teaching a system to *understand* human language in the way we understand language easily leads to a combinatorial explosion of different phrases that the system needs to be able to deal with. For that reason I propose an automation of this process, because using manpower for entering phrases certainly consumes more time and resources than letting an agglomeration of powerful servers crawl the internet to search for phrases and the context in which these phrases appear. Once these phrases have been extracted, they can act as a training set to teach the system. Why can these phrases be considered a good training set? I certainly do not believe that the internet only consists of well-formed phrases. However, the assumption is made that there are more phrases on the web concerning a certain topic that users are able to understand than phrases they cannot understand. Ultimately, the phrases appearing on the web are a result of human language, which itself is mostly clear enough to help people understand the intended meaning. Hence, phrases that can be understood appear more frequently on the web than phrases that cannot be understood. Therefore, the former are represented more frequently. As will be seen later, this assumption forms the basis of automated language extraction in my approach.

It remains an open question, however, whether the extracted information is really enough to assist a system in being able to *understand* language. According to Russell [12] and Chomsky [14], human beings have intuitions regarding the grammaticality of phrases. Chomsky is often called a *nativist* because he assumes that these intuitions result from innate machinery (e.g. Chomsky's notion of a universal grammar). If a universal grammar is indeed a condition to understand language, a system like the one proposed by myself or the one proposed by Doug Lenat and Cycorp would ultimately fail to understand human language, because none of the present servers have the innate machinery of a universal grammar, which itself is probably a consequence of evolutionary processes. Despite recent advances in evolutionary computation [15], web-servers do not undergo the same evolutionary pressures as humans when they acquired the ability to use language as a means of communication. However, Chomsky's theory is not the only one referring to language development. In developmental psycholinguistics, there is an extended debate between different schools of thought. Whilst the Chomskyan position is also referred to as Rationalism, the two other schools of thought Empiricism and Pragmatism provide alternative explanations [12]. Empiricism is considered an example of associative learning, where infants are confronted with speech input. The co-occurrence of specific utterances, words and phrases mediates their learning process. Many connectionist models / artificial neural networks have been proposed to simulate young infants' language learning [16]-[20]. This paradigm would certainly benefit from a large training set via extracting phrases from the internet. Because of a lack of explicit grammar rules, however, the ideas put forward in Empiricism would probably not be sufficient to help an artificial system understand language, e.g. [21]-[24]. The third paradigm, Pragmatism, which is supported by Charles Saunders Peirce, Richard Rorty, Jean Piaget and Michael Tomasello [25]-[31], considers the infant as actively constructing a grammatical inventory by combining knowledge from socio-cognitive interactions with more general learning abilities [12]. This paradigm would also benefit from a large training set via extracting phrases from the internet. Moreover, it would assume an interaction between the servers' present machine learning algorithms and the training set. This interaction may result in a continuous update of existing knowledge by incorporating the ever changing knowledge from the internet. It remains an open question, however, whether present machine learning algorithms would result in the same language competence as the one due to general learning abilities in infants. Moreover, infants have a rich socio-cognitive environment that might help them construct knowledge and refer it to language. This certainly cannot be said for knowledge residing on web servers. Nevertheless, there is at least some hope that an artificial system might learn to understand human language.

Prior to presenting my own approach, I would like to refer to the Wikipedia project, as it partly overlaps with

Doug Lenat's approach. Wikipedia is a widely known online encyclopedia where users can get information about almost anything. This is possible because of a large number of experts all around the world who are willing to make their expert knowledge available to Wikipedia, e.g. an expert on Greek philosophers might provide his/her expert knowledge on Plato, Aristotle or Socrates. An expert on football / soccer might be able to tell about football teams having won the world cup in the past, etc. Obviously, there is the danger that wrong information is put on the web. For that reason, there are other users who can check data entries and correct them in case they spot any wrong information. Since the Wikipedia project is based on volunteers rather than an automated process of extracting information, one might assume that it is relatively slow in updating information. It should be noted that this does not necessarily have to be the case. Seconds after Italy had won the penalty shoot-out against France in the 2006 Fifa World Cup final, Wikipedia not only presented the score of the Cup Final, but also updated Italy's number of World Cup Titles from three to four. Where Wikipedia struggles, however, is when asking a specific question such as "can guinea pigs swim?" In this case, one would require a quick and simple answer such as "Yes" or "No." In order to give this answer, the system would have to understand the question in the first place, which certainly is not possible with the current version of Wikipedia.

II. A NEW APPROACH TO THE SEMANTIC WEB

Having mentioned that a system would have to be able to understand human language in order to permit question-answering in the most efficient way, the approach presented in this section aims to answer how this goal may be achieved. Given that manually teaching a system by providing information such as "water is wet" or "every person has a mother" will soon lead to a combinatorial explosion when more complicated phrases / relative clauses are used. Because this practice requires an almost unlimited amount of resources, the present approach automates this effort. Prior to going into further details, a brief overview will be given. First, the system crawls websites in order to find information. This works in much the same way as the so-called robots of Google and Yahoo crawl websites. Because these robots find the entire text in the sequential order it has been placed on a website, they automatically collect a large number of phrases. It has to be kept in mind, however, that these robots do not interpret a specific set of words as a phrase, nor do Google or Yahoo consider specific strings of words as a phrase. Moreover, they do not even interpret an end of sentence marker as the end of a phrase. Rather, they just collect a large amount of words in sequential order without making any judgment on meaning or grammaticality. Consider the following example: The two phrases "A romantic evening is what girls like. Boys enjoy romantic evenings too, but…" contain the string of words "girls like boys". It should be considered that the end of sentence marker "." as well as capital letters are typically ignored when words are processed by Google or Yahoo. In both phrases, the "girls like boys" sequence contains exactly the same string of words, although the meaning is different. With regard to the task of Google or Yahoo robots, this certainly does not make any difference, as they do not process meaning or grammaticality.

The second step in my approach is to apply an algorithm in order to identify the actual phrases. Technical details of the algorithm will be provided later, as this is meant to be an overview.

The third step consists of storing these phrases in a relational database management system and linking these phrases to possible queries, e.g. when storing the phrase "animals can swim", it can be linked to queries such as "can animals swim?", "do animals swim?" etc. Moreover, the word animal or its plural form "animals" can be linked to phrases such as "a dog is an animal", "a guinea pig is an animal", "a camel is an animal" etc. By referring to this double link, the question "can camels swim?" could be answered correctly even if there was no phrase on the web saying "camels can swim." This is because there are examples on the web saying that camels are animals and that animals can swim. The database stores these examples and links them accordingly. If no answer is stored to a specific query, the query will nevertheless be stored in a database. The purpose of this is to immediately link it to a possible answer once the algorithm identifies it. For the period in between, however, the system will make use of a classical search engine if no answer has been found in the database.

In any case, steps two and three are used to continuously teach the system human language which may ultimately lead to better communication between users sending queries and the system providing answers. It is important to note that the system is updated on a continuous basis. Whilst it accepts queries and searches for related answers in the database, it crawls websites, applies the algorithm to identify phrases and stores new answers / facts in the database, not to mention the new links it forms at the same time. As can be seen, many processes take place in parallel. Flow-charts of these processes are displayed in Fig. 1 and 2. This approach has some reminiscence of both the Semantic Network Theory proposed by Ross Quillian [32] and the theory of neural networks, see [33]-[35] or [36]-[37] for a historical overview, as these approaches also make use of parallel processing. Something to consider too is that the internet changes on a daily basis. Information from today may no longer be present tomorrow. Continuous changes certainly have an influence on phrases that are stored on the database. It is therefore important that the database keeps records of earlier entries that it links to incoming information. This way the database keeps a constant repertoire of knowledge in spite of the fact that the internet is subject to daily changes. New entries and old entries can be linked to each other and more recent entries can be weighted more heavily than older ones. Now it might be asked how these links actually work. Take the following example. A crawler might come across the string "the cat sits on the mat." As mentioned before, it will not regard it as a phrase with a start and an end. However, the algorithm described in this paper will be able to indirectly infer that it is a sequence belonging together. At the time this paper was written, the string "the cat sits on the mat" receives a relatively high number of 62 hits on Google. Including examples from individual websites where other words appear in front of this phrase, such as "like the cat sits on the mat" and running a search on them results in almost no hits. This is because the word like only appears in front of this sequence on one website and other words might appear in front of this sequence on other websites. However, these words vary from website to website. The same is the case for words appearing after the word “mat.”

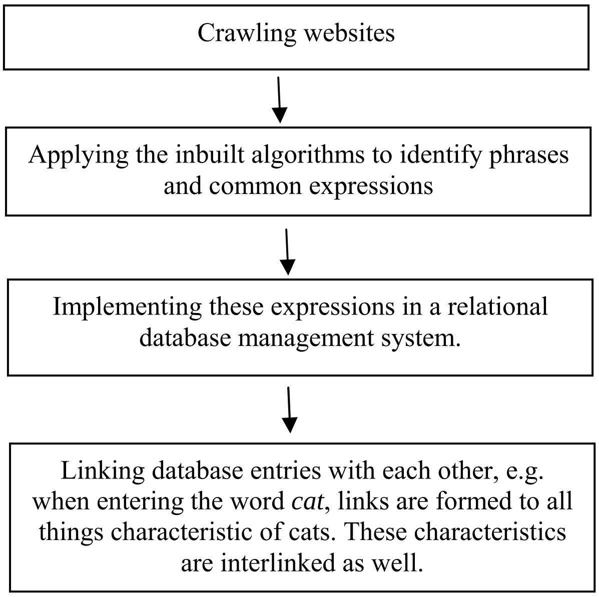
Crawling websites
Applying the inbuilt algorithms to identify phrases and common expressions
Implementing these expressions in a relational database management system.
Linking database entries with each other, e.g. when entering the word *cat*, links are formed to all things characteristic of cats. These characteristics are interlinked as well.

Figure 1. A flow-chart of the system (although the displayed process takes place in a sequential order, many of these sequences run in parallel and are implemented in the database at the same time by making use of a cluster of servers).

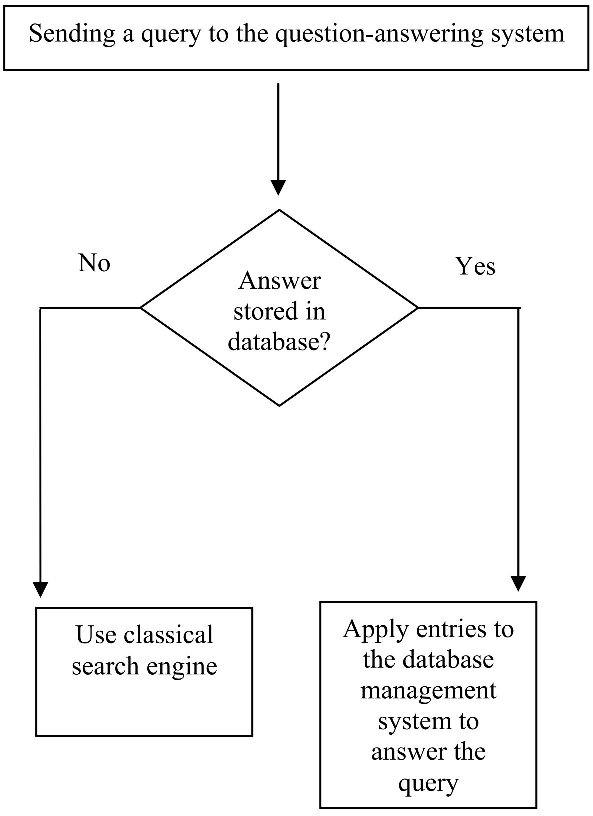
Sending a query to the question-answering system
Answer stored in database?
No
Yes
Use classical search engine
Apply entries to the database management system to answer the query

Figure 2. A flow-chart of a search query (sending queries also runs in parallel to the processes described in Figure 1).

Shortening the phrase will result in a very high number of hits for the expressions “the cat”, “the cat sits on”, “the cat sits on the”, “sits on”, “on the” etc. All these expressions will be linked to other search results containing the same expressions, e.g. “the cat sits on the wall”, “the cat sits on the sidewalk”, “the cat sits on the tree” etc.

This way everything typical of cats is stored in a database (via links). Some things are more typical and therefore there are more examples of it on the web. Likewise, it will be represented through higher activity values in the database. The way these activity values are computed will be described in the section on the system’s algorithm. Not just all things characteristic of cats will be stored in this context, though. The verb “sit” does not apply to cats only. Therefore, links to all living things (people, animals, etc.) who are able to sit will be formed. Even more important, there are many phrases on the web saying “a cat is an animal.” Similarly, there are phrases saying a giraffe is an animal, a dog is an animal etc. Given that there is a link from cat to the word “animal” and because there are links from all other species of animals to the word “animal”, the system will be able to form a category of the word animal. Slowly by slowly, the system is trained to learn a lot of concepts that are common-sense to people, but used to be difficult to understand for machine-learning devices thus far. The internet with its sheer quantity of training examples will be able to continuously train the system 24 hours a day, 7 days a week, 365 days a year. Although it might go too far to assume that machine consciousness will emerge out of this rigorous training, it is realistic to infer that this will ultimately help artificial devices such as search engines to understand human language better than they presently do. Moreover, there are several practical examples where it is advantageous to store information this way. Consider sending the query: “what are the top 5 teams in league X” or “what are the 5 top teams in league X.” Both queries ask for the same content and people have no difficulty to understand the query. However, such a query poses enormous difficulties for present search engines and question-answering systems. By having a fully-interlinked database structure, the links would ultimately link to the same teams. Nevertheless, one caveat remains in order. The league tables representing various sports teams change quickly, as games take place on a frequent basis. This problem is overcome in the following way. Earlier in this paper I had already mentioned that more recent web entries are weighted more strongly than older ones. Given that the most recent tables entirely replace the old ones on the majority of websites (old ones can hardly be found anymore), they end up being weighted much stronger, as the sheer quantity of league tables on the websites of sports channels, newspapers, discussion forums ensures a high activity for the most recent entry.

Apart from these examples, the approach in this paper also comes across a large amount of common-sense knowledge similar to the kind of phrases that Doug Lenat and his colleagues at Cycorp teach their system. These common sense examples will be used to explain the nature of the algorithm. Later, more specialist examples will be discussed in this context as well. Prior to the actual examples, there will be a brief discussion of the theoretical foundations on which the algorithm is based.

III. A SEQUENCE LEARNING ALGORITHM APPLIED TO THE SEMANTIC WEB

The following algorithm was first introduced to simulate research on human learning and memory on a sensory-motor sequence learning task [4], [6], [8]. It is based on evidence from Experimental Psychology, Cognitive Neuroscience and Machine Learning. Although the Semantic Web inherited many features from research into human memory (especially the previously mentioned Semantic Network Theory as well as the numerous papers on biological and artificial neural nets), it came as a surprise that the main equation of this algorithm could be transferred to the Semantic Web approach almost one to one. What additionally inspired the transfer from experimental findings was some evidence of language acquisition in developmental psycholinguistics [12], [16]-[23], [38]-[39]. The infants studied were very young (less than one year old). The main question is how they succeed to learn human language. Moreover, does the inspiration coming from research into language acquisition open new ways to teach artificial systems language learning? My own point of view is that it at least partially does:

First, the infant brain forms an extremely large number of connections during this critical period. This process is modeled by forming links between new and already stored elements. The success of artificial neural networks to simulate language acquisition experiments is another example where concepts are interconnected through weighted links [16]-[20]. Although it has been suggested that artificial neural networks alone would be insufficient to explain the entire process of language acquisition [21]-[24], it has not been doubted that they partially contribute to it (e.g. in a form of hybrid rule-based/associative network). The encouraging finding was that it had been possible to simulate young infants' language learning with a hybrid system as well, where rules emerged out of parallel associative processes [38]-[39]. This gives reason to assume that such a system can also contribute to language learning in the semantic web approach I am describing in this paper.

Secondly, the semantic web approach is trained with a huge training set, merely consisting of text from any website available on the internet. This is a larger training set that any human being could possibly receive in a lifetime. Consequently, the information placed on these websites share a great amount of redundancy, with spelling mistakes included. Similar aspects (though not to that extent) happen with regard to language learning in infants. Once children grow older, they learn to read. In spite of several spelling mistakes that they will certainly come across, humans are nevertheless able to understand the content of most written documents. In software developments (including search engines), these spelling mistakes are coped with by relying on fuzzy logic [40]-[44]. In the semantic web approach discussed here, several fuzzy components are implemented as well (based on [6]), but it is important to note that the web itself contains different forms of spelling, because people make spelling mistakes when implementing websites or blogs. Because the correct spelling is certainly more frequent, however, the approach has no difficulty figuring out the correct spelling and linking to incorrect ways of spelling the same words. As a result, different ways of spelling ultimately link to the same concept. This is a major prerequisite for the understanding of written text. What makes infant language learning different from any semantic web approach is the fact that infants learn language whilst interacting with their social environment, e.g. mums and dads show their kids a teddy bear and say "look at the nice teddy bear" etc. Kids visually perceive the teddy bear, touch it, play with it, listen to the sound of their mum's/dad's voice, try to imitate it etc. [12], [27]-[31], [45]. This is certainly not the way any semantic web learns language. On the other hand, children learn to understand written language much later in life and written language is what the semantic web needs to be able to understand. Common didactics of teaching children written language are based on using a lot of pictorial information (e.g. showing a picture of a car and spelling the word "car" next to it). Because children are pretty proficient in spoken language before they learn to read and write, it can be assumed that their learning of written language is at least partly influenced by their proficiency of spoken language. As a consequence, language learning of semantic web approaches and infants are based on several different processes. Nevertheless, the overlapping features (e.g. a gradual learning of words/phrases/concepts, a network of links between them, a large degree of redundancy with fuzzy elements to be able to deal with spelling mistakes) might help to form an interface between human language proficiency and language understanding of semantic web approaches.

What follows is a concrete example, taken from [5], [45]. A crawler might come across the sequence "men love women." The algorithm of my semantic web approach then puts this string in quotation marks and makes another search on the web. It figures out that there are no consistent, ever-repeating words in front of the word men or behind the word women. Consequently, it regards the expression "men love women" as a closed phrase, especially because its number of hits is enormous (22,900 Google hits at the time search was performed). It then looks whether there is a reciprocal relationship and figures out that the expression "women love men" is also a very frequent one (18,200). If statistics alone were relied on, it would only become obvious that both are very frequent phrases and therefore seem to have a large influence on human language and life more generally. If the reciprocals are added to a total sum of 41,100, the phrase "men love women" would account for 55.7 percent (22,900 out of 41,100), whilst the phrase "women love men" would account for a slightly lower proportion (44.3 percent, i.e. 18,200 out of 41,100). But how does the algorithm actually capture that both are very strong concepts in human language, and how does it assign high activities to these phrases. Because the percentages are similar for both phrases, they should have a similar chance of being activated. Activation depends on a non-linear activation function that was already applied successfully to simulate human learning and memory in sensory-motor sequence learning experiments [4], [6], [8]. Similar to learning and memory models in the cognitive neuroscience community [35]-[37], a concept can be strengthened by excitation. The same concept will also decay over time in case it has not been refreshed. In the present example, excitation increases the activity of a particular phrase and decay decreases it. The more often a particular concept has been excited in the past, however, the less it will decay in the future. This should be clarified with a simple example. Imagine a medical student learning the Latin expressions of anatomy. S/he might learn the name of a particular vein to take blood from. This vein is called “vena mediana cubiti.” In the first few weeks of the course, the name of the vein is easily forgotten. With ongoing repetition of this name (and future training on the way to become a doctor), this name becomes so obvious that it will hardly be forgotten anymore. Even if a person with complete medical training switched directions later on in life, with no contact to medicine for many years, s/he will hardly ever forget the name of this particular vein. Other less common names s/he might forget, e.g. the “ampulla duodeni major”, i.e. the location where the pancreas is connected to the duodenum. Since s/he had learned this concept by heart a long time ago, however, a single repetition might activate it so strongly that it will stay active for a long time again. This is certainly not the same when medical students in their first year hear this term once and decide to drop out of medical school afterwards, because these students missed the chance of repeating this concept many times and therefore it cannot get activated as much as the same concept in a person who completed the entire medical training. Consequently, the activation of a phrase as a concept, its establishment as a memory trace and its decay over time follow highly non-linear processes. These processes were captured in the following equation (1). In this equation, excitation and decay occur separately, as there can be no excitation and decay at the same time. This can be applied to the reciprocals from the example “men love women” / “women love men.” When looking at the sum of both reciprocals (41,100 = 100 percent) and when presenting these 41,100 phrases in random order, each time a phrase is presented there is a probability of *p*=0.557 for the phrase “men love women” and a probability of *p*=0.443 for the phrase “women love men.” Whilst the phrase representing the concept “men love women” is excited, the concept “women love men” decays and vice versa.

$$a_{i+1} = 1 - \left[\frac{1}{1 + \beta \left(e^{\left(\left(\frac{\beta}{1-a_i} \right)^{\frac{1}{2}} \right)} \right)} \right] + \lambda \left[a_i - \frac{a_i}{1 + \eta \left(\frac{1}{\varsigma} \right)} \right] \tag{1}$$

When a phrase is excited in equation (1), β will take on the value 1 and λ will have the value zero. In this case, the excitatory part of the equation (i.e. the part in front of the + sign) is active, whilst the decay part (behind the + sign) remains inactive. During decay, β will have the value 0 and λ will take on the value 1. α_i stands for activity and α_{i+1} represents the next activity value, i.e. the value resulting from either excitation or decay. The exponential’s exponent would be undefined for an α_i value of 1, because its denominator would be zero in this case. Consequently, α_i will be corrected to 0.999 in case it reaches the value 1.

As mentioned in the medical student example, the activity of a concept decays whenever it is not refreshed. Decay of a concept is therefore dependent on the number of its previous activations. If a particular concept has been activated many times in the past, its rate of forgetting will have become slower. The number of previous excitations of a particular concept is denoted by η. Another decay parameter is ς. If this parameter has a low value, decay will be less. If its value is larger, decay will be stronger. Longer phrases that are less likely to appear in exactly this word constellation should therefore have a lower value of ς, for they will be forgotten more slowly in this case. Short phrases are more likely to recur and therefore higher values of ς can be chosen. However, this parameter can be set voluntarily and is not necessarily needed for the success of this algorithm. By default it can be set to 1. Its original purpose in the sensory-motor learning experiments was based on the number of competing sequences a person had to learn, e.g. if a person was trained on eight sequences for 80 trials, each sequence would be presented ten times, i.e. every eighth trial. If a person was trained on four sequences for 80 trials, each sequence would be presented twenty times, i.e. every fourth trial. If a person receives the same sequence every fourth trial, decay should be less than if a person receives the same sequence every eighth trial.

Returning to the example “men love women” / “women love men”, it is important to note that both phrases will end up having a very high activity. This is in spite of the fact that “men love women” are excited in 55.7 percent of the 41,100 presentations and “women love men” are excited in 44.3 percent of the presentations. In other words, the concept “women love men” decays more often than it is excited. Yet this does not mean that it ends up with zero activity. The many excitations from the past have made decay very slowly. In this case, there are thousands of excitations for both examples, so decay becomes almost zero. The same holds true in real life. A professional (e.g. a doctor) who is confronted with essential elements of her/his job on a daily basis will, under normal/healthy conditions be unlikely to ever forget these essentials (decay has reached zero). Fig. 3 displays activity on the Y-axis, whilst the number of training trials appear on the X-axis. For this example, a 50 percent chance of excitation/decay was chosen. As it can be seen, decay becomes less once more excitations have occurred in the past. The example is taken from [6].

Both phrases “men love women” / “women love men” resulted in an equally strong activity of 0.99. In spite of the fact that “men love women” is more frequent than “women love men”, the algorithm has assigned an equally strong activity to both concepts. As it turns out, an approach solely based on statistical information or probabilities would not have achieved this result. A pilot survey carried out by the author on people from different countries, however, yielded a result that was more in line with the algorithm rather than statistics. People in fact regard both phrases to represent very strong concepts in daily life. Therefore, the high activity for both “men love women” and “women love men” seems at least partially justified. In this context it is noteworthy that same-sex relationships are also represented on the web, thus strong activities also exist for “men love men” or for “women love women.”

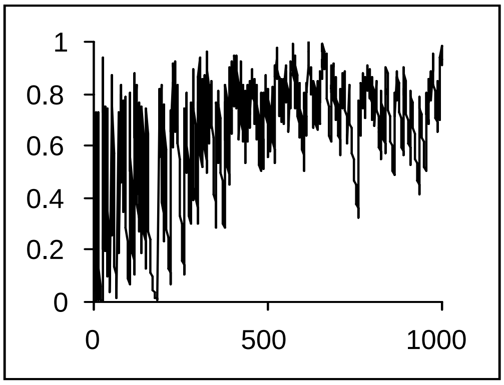

Figure 3. Activity on Y-axis, number of training trials on X-axis. The example is taken from [6]. Upwards direction of the curve represents excitation, downwards direction decay. With an increasing number of excitations, decay becomes less steep.

In the following section an example will be discussed where no reciprocal exists. It is usual that “boys like toys”, but it is hard to imagine a concept for “toys like boys”, though this phrase is certainly possible within a wider context such as “…playing with *toys like boys* anywhere…” [5], [45]. When this system was tested, there were 173 Google hits for “boys like toys” phrases compared with only 3 “toys like boys” examples. Consequently, the total sum of trials this system was trained for was 176. The resulting probability of exciting the phrase “boys like toys” was therefore 98.3 percent, whilst “toys like boys” only had a 1.7 percent excitation probability. Running the algorithm on this problem yielded an activity of 0.98 for the phrase “boys like toys” and 0.001 for the phrase “toys like boys.” When applying this algorithm, there remains a low chance that on trial 176, the opposite phrase “toys like boys” is excited and therefore results in a relatively high activity of 0.73. Such a high activity would certainly contradict the low frequency of the phrase “toys like boys.” In order to get round this problem, computing an average activity over the past 20 or 30 percent of activities gives a more realistic picture.

It might be asked what actually speaks in favor of an algorithm with such strong effects of excitation or decay. Why do excitation and decay not occur gradually? This might be easier understood in the light of the following example. Take the German phrase “München ist Landeshauptstadt des Freistaates Bayern” (Munich is the capital of the Free State of Bavaria). Appearing on only three websites, this example produced a very rare number of hits in Google. Since there are no competing hits, this information is nevertheless likely to be correct (e.g. entering “ist Landeshauptstadt des Freistaates Bayern” produced no hits with alternative cities in front of the string “ist Landeshauptstadt des Freistaates Bayern”). Why should a piece of information that is likely to be correct not reach a high activity. This is exactly what the previously mentioned algorithm makes possible. If activity increases fast and there is no decay due to the absence of competing information, activity reaches high values quickly and stays there. This is vital for rare information, e.g. long and complicated expressions that cannot be found on the web so easily. By having 3 consecutive excitations throughout the 3 training trials, it is possible to get an activity as high as 0.95 for “München ist Landeshauptstadt des Freistaates Bayern.” It is worth comparing this example with the impossible “toys like boys” expression, as both had exactly 3 hits in Google. “Toys like boys” ended up having such a low activity due to its competing reciprocal. The “München ist Landeshauptstadt des Freistaates Bayern” reached such a high activity because there were no competing cities and there were no reciprocals to compete with, as “Landeshauptstadt des Freistaates Bayern ist München” produced no hits in Google. Although the reciprocal from the previous phrase is not grammatically wrong, it is not common to express information about a city in this text order. When saying that a particular city is capital of a particular state, the city appears first in the overwhelming majority of phrases. These are ways how this approach indirectly captures concepts and expressions in a given language.

IV. CONCLUSIONS

Having discussed the way the algorithm works, it will be summarized in the light of a new example. This example will also be used to show how it interacts with the relational database management system. Finally, practical applications are considered. Imagine you have a specific question and aim to receive a quick answer to this question. You may have two guinea pigs and you wonder whether they are able to swim. You would ask the system: “can guinea pigs swim?” Present search engines will not be able to give you an answer by saying “Yes” or “No”, because they do not understand the question in the first place. The semantic web approach discussed here has different ways of answering this question, though. First, its crawler might have come across a phrase like “guinea pigs can swim” in the past. Therefore, this phrase might be stored in the database already. Because the database also has a connection between the terms “can” and “yes” and “cannot” and “no” and because there are many websites saying that “guinea pigs can swim” with essentially no websites saying “guinea pigs cannot swim” / “guinea pigs can’t swim”, it can not only answer the question by saying “guinea pigs can swim” or by simply saying “yes.” This is because the system has learned that questions in the form “can…?” require an answer such as “yes” or “no.” This is not the only way how this semantic web approach can answer the question, however. Because there are websites saying “the guinea pig is an animal” and other websites saying “all animals can swim”, the database is able to form a link between “guinea pig” and “animal” and between “animal” and “can swim.” Hence, asking “can guinea pigs swim” is able to elicit the same answer. When asking about dogs, cats, tigers, zebras, elephants etc. the answer would therefore be the same. Even when asking about an animal where no website contains text about swimming, this approach will be able to infer the answer. For instance there is no website saying “a lama can swim”, but there are websites saying that it is an animal and other websites saying that all animals can swim. Due to the database-links between these concepts, this semantic web approach copes with the problem.

On the other hand, there is a danger of loosing too much information, though. There are websites providing more details, such as the fact that guinea pigs can swim just like all animals can swim by instinct. These websites also say, however, that guinea pigs can only swim for a short time and people who put them into water will risk them to die from a heart attack. If one relied on the short answer "yes" after asking whether they can swim, one might be tempted to put them in the bathtub and end up making the sad experience that they do not survive the bath. The same might be true for many animals. Keep in mind that humans can swim by instinct too, but it far too often happens that people drown. So reducing the information to a very short "yes" or "no" might not always be advantageous, much as in daily life. Nevertheless, a quick answer might be good to get a first orientation in a confusing situation. Such a semantic web approach would therefore also have great potential for the mobile phone industry. A question could be sent as a text message and an answer could be collected almost instantly. Alternatively, the question could be spoken, recognized by a speech recognition device and the answer could be given instantly (either in text format or after transformation into speech). Even should Noam Chomsky's idea on a universal grammar be correct and it should turn out that computers cannot develop this type of universal grammar because it has originated from evolution, there might be a lot of practical benefits resulting from this approach. In spite of completely understanding human language, the system might emulate this process. Although an emulation is not the same process in nature, it may succeed to enable sophisticated question-answering by artificial systems. Finally, devices of this kind might ultimately change society. Rather than memorizing a lot of facts, it would be more important to memorize strategies in order to find out about these facts. If almost all knowledge of the world can be accessed instantly and almost anywhere by more sophisticated search, things we are taught in school might become different too. Our grandparents' generation did not own electronic calculators or personal computers at home, so their mathematical calculations had to be much more primitive than nowadays. Similarly, young doctors back in the old days learned essentially no biochemistry in medical school, let alone any recent techniques in molecular biology that already impact on today's therapy and diagnostic methods.

Pointing to the hypothesis that tomorrow's semantic web technologies might be able to cover almost all knowledge of the world, in a Tower of Babel manner, one might be tempted to say that we have been progressing faster to this scenario with recent developments in the semantic web community, e.g. [46]-[55]. Take Google's Print Project as an example, where Google has teamed up with the New York Public Library and the university libraries of Oxford, Stanford, Harvard and Michigan in order to crawl the text from every book and to make this text available in public. It should be kept in mind that Oxford University library carries a copy of every book that has ever been printed in Britain. This is an enormous number, not to mention the increase of the number that will result from adding all books from Widener library (i.e. the world's largest university library, located at Harvard) and the New York Public Library. In France it is also considered to make the written knowledge from public libraries available online. This will suddenly boost the amount of publicly available information written in French language. The result of making this information available is likely to be an improvement of question-answering systems such as the one described in this paper. Implementing this knowledge in a linked database will certainly not result in adverse effects when trying to help an artificial system understand human language.

ACKNOWLEDGMENT

R.S. thanks Wolfson College, University of Cambridge (UK) for electing him a Research Fellow, which enabled him to develop this semantic web approach. He also thanks the Technology Transfer Office, University of Cambridge (UK) in association with the partnership (CMI) between the University of Cambridge (UK) and the Massachusetts Institute of Technology (MIT, USA) for opening the case file to promote his research on this topic.

REFERENCES

[1] D. Fensel, "Ontology based knowledge management," *IEEE Computer*, vol. 35, pp. 56-59, November 2002.

[2] N. Cercone, L. Hou, V. Keselj, A. An, K. Naruedomkul and X. Hu, "From computational intelligence to web intelligence," *IEEE Computer*, vol. 35, pp. 72-76, November 2002.

[3] S. Sekine and R. Grishman, "Hindi-English cross-lingual question-answering system," *ACM Transactions on Asian Language Information Processing*, vol. 2, pp. 181-192, September 2003.

[4] R. Spiegel, "Human and machine learning of spatio-temporal sequences: an experimental and computational investigation," PhD-thesis, University of Cambridge (UK), 2002.

[5] R. Spiegel, "A machine learning approach towards improving internet search with a question-answering system," in *4th Int. Conf. on Computational Intelligence, Man-Machine Interaction and Cybernetics*, L. Zadeh and K. Grigoriadis, Eds. Miami: World Sc. Eng. Acad. and Soc., 2005, pp. 270-275.

[6] R. Spiegel, M.E. Le Pelley, M. Suret and I.P.L. McLaren, "Combining fuzzy rules and a neural network in an adaptive system," *Proc. IEEE Int. Conf. Fuzzy Systems* (IEEE World Congr. on Comptl. Intelligence), 2002, pp. 270-275.

[7] R. Spiegel and I.P.L. McLaren, "Abstract and associatively-based representations in human sequence learning," *Phil. Trans. Roy. Soc. London*, vol. B358, pp. 1277-1283, July 2003.

[8] R. Spiegel and I.P.L. McLaren, "A hybrid cognitive-associative model to simulate human learning in the serial-reaction time paradigm," *Proc. AISB: Adaption in Artificial and Biological Systems*, T. Kovacs and J.A.R. Marshall, Eds. Bristol: Society for the Study of Artificial Intelligence and Simulation of Behaviour, vol. 1, pp. 74-90, April 2006.

[9] R. Spiegel and I.P.L. McLaren, "Associative sequence learning,“ *J. Expt. Psychology: Animal Behavior Processes*, vol. 32, pp. 150-163, April 2006.

[10] L. Zadeh, "From search engines to question answering systems – the problems of world knowledge, relevance and deduction. Keynote lecture given at the WSEAS Fuzzy Systems Conference, June, 2005.

[11] M. Leslie, online publication on Doug Lenat and his Cycorp team. http://www.stanfordalumni.org/news/magazine/2002/marapr/departments/brightideas.html

[12] J. Russell, *What is Language Development? Rationalist, Empiricist, and Pragmatist Approaches to the Acquisition of Syntax*, Oxford: Oxford University Press, 2004, pp. 1-11.

[13] N. Chomsky, *New Horizons in the Study of Language and Mind*, Cambridge, UK: Cambridge University Press, 2000, Foreword p. XIV.

[14] N. Chomsky, *Rules and Representations*, Oxford: Basil Blackwell, 1980, pp. 189-192.

[15] A. Gosh and A. Tsutsui, *Advances in Evolutionary Computing*, Berlin: Springer, 2002.

[16] E. Bates and J.L. Elman, "Learning rediscovered," *Science*, vol. 274, pp. 1849-1850, December 1996.

[17] J.L. Elman, E. Bates, M.H. Johnson, A. Karmiloff-Smith, D. Parisi and K. Plunkett, *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT-Press, 1996.

[18] K. Plunkett and J.L. Elman, *Exercises in Rethinking Innateness*. Cambridge, MA: MIT-Press, 1997.

[19] P. McLeod, K. Plunkett and E.T. Rolls, *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press, 1998.

[20] M. Negishi, "Do children learn grammar with algebra or statistics?" *Science*, vol. 284, p. 433, April 1999.

[21] G.F. Marcus, S. Vijayan, S. Bandi Rao and P.M. Vishton, "Rule learning in seven-month-old infants," *Science*, vol. 283, pp. 77-80, January 1999.

[22] S. Pinker, "Out of the minds of babes," *Science*, vol. 283, pp. 40-41, January 1999.

[23] G.F. Marcus, *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT-Press, 2001.

[24] G.F. Marcus, "Can connectionism save constructivism.," *Cognition*, vol. 66, pp. 153-182, May 1998.

[25] C.S. Peirce, "The icon, the index and the symbol," in *Collected Papers of Charles Saunders Peirce*, vol. II, C. Hartshorn and P. Weiss, Eds. Cambridge, MA: Harvard University Press, 1932, pp. 156-173.

[26] R. Rorty, "Universality and truth," in *Rorty and his Critics*, R.B. Brandom, Ed. Oxford: Basil Blackwell, 2000, pp. 1-24.

[27] J. Piaget, *The Psychology of Intelligence*. London: Routledge and Kegan Paul, 1948.

[28] J. Piaget, *The Child's Conception of Number*. London: Routledge and Kegan Paul, 1952.

[29] M. Tomasello, *First Verbs: A Case Study in Early Grammatical Development*. Cambridge UK: Cambridge University Press, 1992.

[30] M. Tomasello, *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah N.J.: Erlbaum, 1998.

[31] M. Tomasello, The *Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press, 1999.

[32] R. Quillian, "Word concepts: a theory and simulation of some basic semantic capabilities," *Behavioral Science*, vol. 12, pp. 410-430, September 1967.

[33] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning representations by backpropagating errors," *Nature*, vol. 323, pp. 533-536, October 1986.

[34] D.E. Rumelhart and J.L. McClelland, *Parallel Distributed Processing*, vol. I, Cambridge, MA: MIT-Press, 1986.

[35] J.L. McClelland and D.E. Rumelhart, *Parallel Distributed Processing*, vol. II, Cambridge, MA: MIT-Press, 1986.

[36] S. Grossberg, *Neural Networks and Natural Intelligence*, Cambridge, MA: MIT-Press, 1988.

[37] J.A. Anderson and E. Rosenfeld, *Talking Nets*, Cambridge, MA: MIT-Press, 1998.

[38] R. Spiegel, "Cognitive modeling of symbolic-like relationships with the adaptive neural network associator (ANNA)," *Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN)*, 2003, vol. 4, pp. 2746-2751.

[39] R. Spiegel, "A novel approach to extract rules from sequences of phonemes," *Proc. Cambridge First Postgr. Conf. on Language Research (CamLing)*, 2003, pp. 494-500.

[40] L.A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-353, August 1965.

[41] L.A. Zadeh, "Fuzzy algorithms," *Information and Control*, vol. 12, pp. 94-102, December 1968.

[42] L.A. Zadeh, "Making computers think like people," *IEEE Spectrum*, pp. 94-102, August 1984.

[43] R. Kruse, J. Gebhardt and F. Klawon, *Foundations of Fuzzy Systems*, Chichester: Wiley, 1994.

[44] V. Loia, M. Nikravesh and L.A. Zadeh, *Fuzzy Logic and the Internet*, Berlin: Springer, 2004.

[45] R. Spiegel, "A search engine inspired by natural language and human memory," *WSEAS Trans. Information Science and Applications*, vol. 3, pp. 56-63, January 2006.

[46] H. Lausen, Y. Ding, M. Stollberg, D. Fensel, R. Lara and S. Han, "Semantic web portals: state-of-the-art survey," *Journal of Knowledge Management*, vol. 9, pp. 40-49, October 2005.

[47] N. Zhong, J. Liu and Y. Yao, "In search of the wisdom web," *IEEE Computer*, vol. 35, pp. 27-31, November 2002.

[48] M.C. Daconta, L.J. Obrst and K.T. Smith, *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management*, London: Wiley, 2003.

[49] V. Alexiev, M. Breu, J. de Bruijn, D. Fensel, R. Lara and H. Lausen, *Information Integration with Ontologies*, London: Wiley, 2005.

[50] B. Leuf, *The Semantic Web. Crafting Infrastructure for Agents*, London: Wiley, 2006.

[51] S. Staab and R. Studer, *Handbook of Ontologies*. Heidelberg: Springer, 2004.

[52] G. Antoniou and F. van Harmelen, *A Semantic Web Primer*. Cambridge, MA: MIT-Press, 2004.

[53] T. Berners-Lee, J. Hendler and O. Lassila, "The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, pp. 34-43, May 2001.

[54] J. Davies, D. Fensel and F. van Harmelen, *Towards the Semantic Web: Ontology-driven Knowledge Management*, London: Wiley, 2003.

[55] D. Fensel, W. Wahlster, H. Lieberman and J. Hendler, *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential*. Cambridge, MA: MIT-Press, 2003.

AUTHOR

R. Spiegel is with the Institute of Medical Psychology, Ludwig-Maximilians University (LMU), Goethestr. 31/1, D-80336 Munich and with the Medical Schools of both TU and LMU in Munich, Germany (e-mail: rainer.spiegel@campus.lmu.de). From 2002 to 2006 he has been Fellow of Wolfson College, University of Cambridge, Cambridge CB3 9BB, United Kingdom.

Manuscript received 15 May 2006.