A Natural Language Processing for Semantic Web Services

Mladen Stanojević, *Member, IEEE*, Sanja Vraneš, *Member, IEEE*

Abstract **— The problem of Natural Language Understanding is one of the first problems researchers in AI were trying to solve, and our brain is the best proof that the problem can be solved. In this paper we propose a model that could be used to describe roughly the process of understanding as it happens in our brain. Using the proposed model we have developed a Hierarchical Semantic Form for the representation of semantics and the corresponding SOUL (Space of Universal Links) learning algorithm. To proof the feasibility of the concept, we have implemented a prototype Semantic Web Service, that uses natural language processing to interpret the conventional, natural language content of traditional Web page and to retrieve the information asked for in natural language as well.**

Keywords **— Natural Language Processing, Knowledge Representation, Artificial Intelligence, Semantic Grammars, Semantic Web, Web Services.**

I. Introduction

SEMANTIC Web is an emerging concept that advocates encoding of information and services in a structured machine interpretable manner [1]. With the promising Semantic Web technologies, it is possible that not only humans but also machines can interpret and retrieve the required information and services. Even with the benefits of Semantic Web technologies, there are no adequate tools or applications, which reveal the potentials of semantically, encoded knowledge and services over Semantic Web. Semantic Web relies today on the use of various ontology and schema languages to represent the semantics of Web pages. However, this requires the translation of a large number of existing Web pages and use of annotated pages which can be avoided if some form of rudimentary Natural Language Understanding is provided.

Researchers in AI invested a lot of effort in an attempt to solve the problem of Natural Language Understanding. The task of understanding natural language input was decomposed into three subtasks:

- Representation of knowledge about natural language (natural language grammars)
- Representing the meaning (semantics)
- Transformation of natural language input (natural language parsing) using grammar knowledge into semantic knowledge representation

Grammars are used to define the structure of natural languages. Since the first papers of Noam Chomsky on linguistic theory and grammars [2] many classifications and grammars have been proposed.

Although some attempts have been made to introduce semantic grammars, syntactic grammars are predominantly used nowadays. This is probably due to the deceptive possibility that relatively simple, general syntactic grammar, able to parse any sentence in natural language, is within our reach. However, even if such grammar is proposed one day, it will not contribute a lot to Natural Language Understanding, because sentences that share the same syntactic structure may have completely different meanings.

The meaning is represented using different techniques in AI [3] and Semantic Web: logic formalism, semantic nets, conceptual dependencies, frames (schemas), scripts, rules, or their combinations [4], as well as different ontology and schema languages such as XOL, SHOE, OML, RDFS, DAML+OIL and OWL.

Of the three subtasks, parsing [5] was the best understood. Parsing is a process of "delinearization of sentences" using knowledge about syntax, semantics, etc, whose task is to give the meaning to a sentence by finding the function of words within the sentence and creating a structure such as parse tree.

On the other hand neurologists claim that our brain is always using the same formalism, hierarchically organized neural network, to represent both, grammar rules and semantics, as well at to perform natural language parsing. Although the role of Sensory Information Storage (SIS), Short-Term Memory (STM) and Long-Term Memory (LTM) is different in the process of Natural Language Understanding, the formalism lying behind each of them is the same.

II. Process of Understanding

Cognitive psychologists [6] came to a conclusion that our memory is not a single, simple function, but a complex system made of diverse components and processes. The three most important components are Sensory Information Storage (SIS), Short-Term Memory (STM), and Long-Term Memory (LTM). Each of them has its own function, capacity, the form of information held, and the length of time information is retained.

This work has been partly supported by the Ministry of Science and Environmental Protection of Republic of Serbia (Pr. No.: TR-6103A).

M. Stanojević is with the Mihailo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia & Montenegro (Phone: +381 11 2774 024; Fax: +381 11 2776 583; e-mail: mladen@lab200.imp.bg.ac.yu).

S. Vraneš is with the Mihailo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia & Montenegro (Phone: +381 11 2772 876; Fax: +381 11 2776 583; e-mail: mladen@lab200.imp.bg.ac.yu).

SIS is responsible for holding sensory images for several tenths of second after they have been received by our sensory organs. Its primary function is to make temporary snapshots which will give enough time to our brain to process it.

Information from SIS is transferred into short-term memory, where it is held again for only a short period of time (a few seconds or minutes). Whereas SIS holds the complete image, STM stores only the interpretation of the image. When we hear a sentence, SIS retains the sounds, while STM holds the words and more complex semantic categories.

A central characteristic of STM is the severe limitation on its capacity. A person who is asked to listen to and repeat a series of 10 or 20 names or numbers normally retains only five or six items. However, these items can represent arbitrarily complex, abstract hierarchies. I am not able to reproduce the 28-digit number, but if I recognize that the first 8 digits represent my birthdate, the next 10 digits – my mobile phone number, and the last 10 digits – mobile phone number of my wife, then I can do it. In my short term memory are not 28 digits, but the three items representing my birthdate, my mobile phone number and mobile phone number of my wife.

We can reason only on the information stored in our STM. The understanding of what we see or hear is stored in our STM.

The basic function of LTM is to observe, create and store patterns and to record all our experiences in the form of these patterns. It is believed that the capacity of LTM is almost unlimited and that it is not limited in time.

Though the process of understanding is still not understood completely, it could be described coarsely as following: the spoken sentence is stored in SIS, compared with patterns in LTM and recognized complex semantic categories are then stored in STM. As soon as semantic categories has been stored in STM, we understood the sentence and can reason further about it, or store it in LTM if we assess that this information is important for us.

To illustrate better the process of understanding imagine that we entered in a tourist agency and asked the clerk:

"Could you please give me the first five morning Lufthansa flights departing from Berlin and arriving at Rome on next Sunday?"

Based on his/her previous experience and training, after the information is stored in SIS and compared with the existing patterns in LTM, the following semantic categories could be extracted in STM:

```
<polite-query> = "Could you please give me"
<flight-time-list > = "the first five
morning Lufthansa flights"
<departure-phrase> = "departing from
Berlin"
<arrival-phrase> = "arriving at Rome"
<time> = "on next Sunday"
```

The brain will probably reason out that <polite-query> semantic category doesn't contain any useful information, therefore it need not be stored in STM. This way, our brain is able to filter out parts of sentences that carry no useful information or we do not understand, which enhances in a great deal our ability to comprehend the meaning of the sentence.

Using the last four semantic categories and knowing flight retrieval procedure, the clerk will be able to find the needed flights for us.

The described understanding process is in contradiction with the belief underlying many natural language grammars, that the whole sentence must be matched with the corresponding grammar rule to be able to understand the meaning of the sentence. Actually, the meaning is not in the whole sentence, but in the main semantic categories that constitutes this sentence.

Another common belief is that syntax of a sentence is very important for understanding. However, the patterns in LTM correspond to semantic categories, not sentences, hence only the syntax of semantic categories is important. This is a trick our brain is using to decrease the number of patterns stored in our LTM. Instead of remembering the syntax of thousands of possible forms for flight information queries, LTM should store the syntax for less than 30 semantic categories. This also provides a great flexibility in understanding of sentences, because understanding is not dependant on the sequence of semantic categories, or on appearances of unneeded information or unknown words or phrases.

To be able to represent the syntax of complex semantic categories, LTM must use some form of hierarchically organized sequences (e.g. <flight-time-list > Fig. 1).

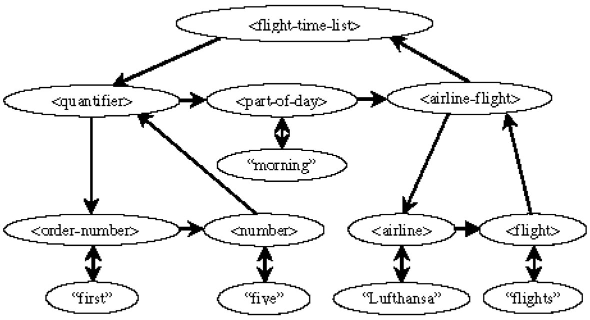


Fig. 1. Complex semantic category

III. Hierarchical Semantic Form

A knowledge representation technique that would simulate LTM should provide means for the efficient representation of syntax and semantic categories and their contexts (links – relationships between these semantic categories). If we use plain text to describe the syntax and semantics, then we can identify letters as the lowest categories, then groups of letters, words, groups of words, sentences, etc. Each of these categories must be represented only once to facilitate creation of new relationships and search of existing relationships. These categories are in their essence sequences. Words are sequences of letters, groups of words are sequences of words, while sentences are sequences of groups of words. As we can see, an appropriate knowledge representation technique should be able to represent the hierarchically organized sequences. The need for hierarchical representation and automatic linking was also recognized before.

The Hierarchical Semantic Form (HSF) [7], which resembles hierarchically organized neural network, can be used to represent various kinds of syntax and semantic categories as well as relationships between these syntax and semantic categories.

The automatic extraction of semantic categories and relations between them is provided by the SOUL (Space Of Universal Links) algorithm, which gives support to the Hierarchical Semantic Form. HSF uses two data types, *groups* and *links*, to build the hierarchy of categories.

The group data type designates uniquely characters, a group of characters, words, a group of words, sentences, etc, while the link data type enables the creation of sequences at different hierarchy levels (sequences of characters, words, group of words, sentences, etc.). The main role of links is to represent categories (groups) in different contexts

Fig. 2 gives one possible graphical representation of HSF, analog to the hierarchy in Fig. 1, for an instance of <flight-time-list> semantic category: "first five morning Lufthansa flights".

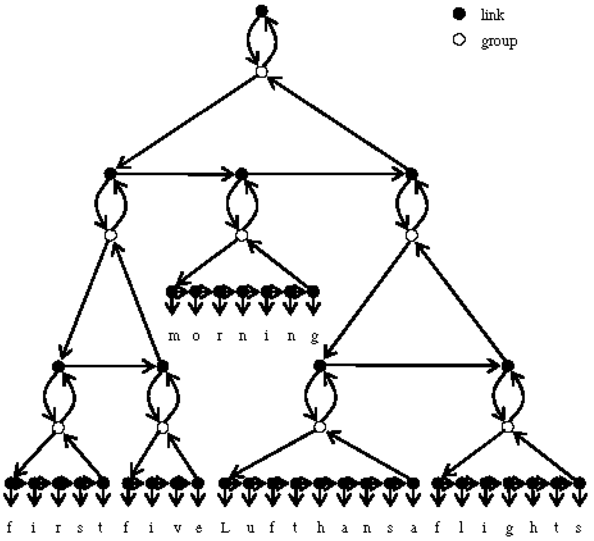


Fig. 2. Hierarchical Semantic Form

The SOUL algorithm builds the hierarchy of groups and links as characters are fed to it. This hierarchy changes dynamically as new characters are processed by SOUL. The HSF in Fig. 2 will be created if first words, "first", "five", "morning", "Lufthansa" and "flights" are fed to SOUL, then groups of words representing higher semantic categories, "first five" and "Lufthansa flights", and finally "first five morning Lufthansa flights".

SOUL algorithm provides an LL(k) bottom-up parser that is able to parse partially the input text and extract complex semantic categories. As a result it provides parse trees of matched semantic categories, which can be used for further processing. These parse trees play the same role as STM in human brain.

Although the capabilities of HSF with SOUL in Natural Language Processing might be similar to the systems relying on "controlled English" [8], there are actually some substantial differences. The systems relying on "controlled English" use syntactic grammars and they are not able to recognize semantic categories and relations between them, but only syntactic categories. The system itself does not understand anything, and programmer is the one who actually gives the meaning to syntactic categories. Furthermore, as the name "controlled English" suggests, these systems are limited to a subset of natural language and do not have a potential to provide full natural language understanding. On the other hand HSF with SOUL is scalable and facilitates easy defining of new semantic categories, thus enhancing cognitive capabilities of the system.

IV. Flight Search Example

ARPA launched Spoken Language Systems program in 1988, a five-year program centered around a pseudo-application called the Air Travel Information Service (ATIS). Inspired with ATIS project, we have developed a prototype Flight Information Service (FIS), a Semantic Web Service, which should provide information about airline timetables. We have defined first meta knowledge, i.e. the definitions of semantic categories and relations used in this domain, and then we defined a flight base (using an ordinary HTML file) in natural language for major European airlines.

As a backbone of Flight Information Service we used SOUL algorithm with HSF, which provide a natural language parser and a knowledge representation technique. We used SOUL Commander, a kind of Natural Language Processing shell, to define semantic categories and relations between them for flight definitions and queries. The learning process starts with learning words ("flight", "Berlin", "morning", "departing", …).

The next stage is learning simple semantic categories (e.g., that Lufthansa, AlItalia and Air France are instances of the semantic category <airline>), while the final stage is defining complex semantic categories comprised of simpler ones already learned. The complete list of definitions can be entered for less than half an hour. Though this list certainly doesn't cover all possible forms of queries, new semantic categories can be easily added to cover these cases.

After semantic categories and relations have been defined, a general form of query and flight definition (comprised of these semantic categories) are defined using SOUL Commander.

These two commands cover thousands of possible query and flight definition forms. Of course, these two forms don't cover all possible cases for queries and flight definitions, but by adding new semantic categories, understanding capabilities can be enhanced.

After the semantic categories and relations have been defined, a flight base is defined in an ordinary Web page in the following form:

"Welcome aboard on a Lufthansa Berlin-Paris flight LH4310 departing at 5:25 in the afternoon and arriving at 7:05 PM each Sunday."

FIS incorporates SOUL Commander, all grammar rules, flight base and stores the context of dialog with user. FIS is storing the context which is deep enough to find all basic information required for a flight query, departure and arrival city and day (date) of flight.

A user can communicate with FIS in a form of guided dialog, but also in a form of complex queries such as:

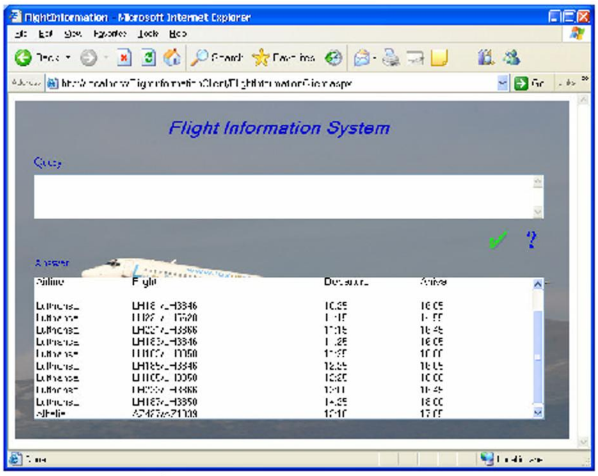
"What flights are there from Berlin to Rome departing after 11:00 o'clock in the morning and arriving between 2 PM and 15:00 on May the 15th?"

FIS will find only one flight matching all the constraints, a Lufthansa flight, LH221/LH5628, departing at 11:15 and arriving at 14:55.

If we would like to see some more flights, we could relax a bit constraint on departure and arrival time by typing:

"First ten flights departing after 10:00 and arriving between 2 PM and 18:00"

and FIS will now present only the first ten flights (Fig. 3), taking into account new constraints and context information about departure city (Berlin), arrival city (Rome) and flight date (May the 15th).


FlightInformation - Microsoft Internet Explorer
Flight Information System
Query
Answer

Airline	Flight	Departure	Arrival
Lufthansa	LH18?/LH3346	10:25	16:05
Lufthansa	LH221/LH5628	11:15	14:55
Lufthansa	LH22?/LH3366	11:15	16:45
Lufthansa	LH183/LH3346	11:25	16:05
Lufthansa	LH10?/LH3350	11:55	18:00
Lufthansa	LH185/LH3346	12:25	16:05
Lufthansa	LH105/LH3350	12:25	18:00
Lufthansa	LH22?/LH3366	13:11	16:45
Lufthansa	LH187/LH3350	14:25	18:00
Alitalia	AZ407/AZ1009	10:10	17:05

Fig. 3. Partial query

Note that FIS could be relatively easy upgraded to TIS (Travel Information Service) by defining few semantic categories, few new commands and timetables for bus and train trips.

V. Conclusions

To allow computer programs to understand natural language was the long-term goal for many AI researchers, which could not be reached mainly because impropriate knowledge representation techniques were used and because partial approach to the problem of Natural Language Understanding was applied. Namely, this complex problem was decomposed into three subtasks: defining natural language grammar, representing the meaning (semantics), parsing natural language input, which were then being solved more or less independently.

However, the neurology and cognitive sciences suggest another approach, a holistic approach, where the same formalism, hierarchically organized neural network, is used to define grammar rules, represent the meaning and parse natural language input, and is underlying Sensory Information Storage, Short-Term Memory and Long-Term Memory, which take part in the process of Natural Language Understanding as it happens in our brain.

Based on possible (not necessarily true) model of understanding natural language in human brain presented in this paper, we have proposed Hierarchic Semantic Form (HSF), which resembles hierarchically organized neural network and represents a hierarchical equivalent of plain text forms, where all semantic categories are explicitly represented and hierarchically organized.

To validate the ideas of HSF and SOUL we have developed a prototype of Semantic Web Service for flight search. Flight Information Service (FIS) shows great flexibility in understanding various flight definitions and queries in natural language. FIS is scalable, because new semantic categories can be easily added to enhance its understanding capabilities and new airlines and new flights can be also defined to increase usefulness of service, robust, because it is not confused by the unknown words and phrases or by syntactically incorrect queries, and portable, because semantic categories used in FIS, could be reused, for example, in an information service about bus or train trips.

References

[1] R. Guha, Rob McCool, Eric Miller, "Semantic Search", in *Proceedings of the 12th international conference on World Wide Web*, Budapest, 2003, pp. 700-709.

[2] N. Chomsky, "On certain formal properties of grammars", *Information and Control*, No. 2, 1959, pp. 137-167.

[3] J. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole Publishing Co., Pacific Grove, CA, 2000.

[4] S. Vraneš, M. Stanojević, "Integrating Multiple Paradigms within the Blackboard Framework", *IEEE Transactions on Software Engineering*, Vol. 21, No. 3, 1995, pp. 244-262.

[5] A. Barr, E. Feigenbaum, *The Handbook of Artificial Intelligence (Vol. 1)*, Pitman, 1983.

[6] R.J. Heuer, Jr, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, 1999.

[7] M. Stanojević, S. Vraneš, "Semantic Web with SOUL", *Proceedings of the IADIS International Conference e-commerce 2004*. Lisbon, Portugal, 2004, pp. 123-130.

[8] Stefan Hoefler, "The Syntax of Attempto Controlled English: An Abstract Grammar for ACE 4.0", Technical Report ifi-2004.03, Department of Informatics, University of Zurich, 2004.