NATURAL LANGUAGE SEMANTICS
AND PROBLEM OF LAYERS

Roman Mouček

Department of Computer Science and Engineering
University of West Bohemia
Plzeň, CZECH REPUBLIC

E-mail: `moucek@kiv.zcu.cz`

Abstract: This paper deals with one of the essential problems connected with processing of natural language semantics – an effort to divide natural language semantics into several layers and process them separately. We introduce some existing solutions according to field of science dealing with natural language processing. Then we discuss the advantages and disadvantages of this approach in contrast to processing of semantics as the indivisible unit. As a result we describe a contemporary approach of our research group to the problem of natural language processing.

Keywords: Natural language semantics, Semantic layer, Neuroscience, Human brain, Linguistics, Artificial intelligence, Dialogue system

1. INTRODUCTION

The computerized processing of natural language semantics, a model of real word concepts and process of understanding to spoken or written utterance belong to the main unsolved tasks of informatics. This problem is partly considered e.g. in applications of artificial intelligence like human-computer dialogue systems or in applications of text classification. The most impressive idea seems to be an effort to realize a semantic web, which should replace the nowadays approach to information search.

However, all these ideas and application prototypes suffer from the enormous complexity and ambiguity of natural language semantics. As the results, the existing applications rely more on processing of morphology and syntax and take into account only a narrow part of reality. An effort to build a general ontology and to divide the processing of semantics into separate tasks faces to a lot of important problems and no model introduced till now can be considered as the suitable one.

The main goal of this paper is to introduce one of the essential problems connected with processing of natural language semantics – an effort to divide natural language semantics into several layers and process them separately. We introduce some existing solutions according to field of science dealing with natural language processing. Then we discuss the advantages and disadvantages of this approach in contrast to processing of semantics as the indivisible unit. As a result we describe a contemporary approach of our research group to the problem of natural language processing.

2. LAYERS OF SEMANTICS

There is no doubt that to process a natural language semantics means to process a complex system, which is closely connected with human thinking and reasoning. The large complexity of this system and our limitations in understanding to this system naturally lead to attempts to identify a set of layers in semantics, which could be processed separately to some extent. The identification of layers depends on the point of view of different fields dealing with problem of semantic processing.

2.1 Cognitive science

The human brain is result of hundred million years of evolution. The complexity of present-day brain is connected with the ability of evolution (evolvability) (Koukolík, 2003). Evolvability can be observed already on the cellular level. Each cell solves the problems connected e.g. with cellular division or cellular metabolism with the help of many subsystems which cooperate to some extent, and which are also separated with semi-permeable membranes.

Taking into consideration the cellular processes, it is supposed that also brains cannot solve all the problems with the help of one large neural network. Our brains are considered to have particular subsystems to solve particular problems. The neurons practically always operate as functional units which can be imagined as two-dimensional maps although they work, ofcourse, as complex three-dimensional systems.

A collection of functional maps is called the brain functional system. Then the human brain can be considered as a collection of widely defined functional systems, e.g. visual system, auditory system, variety of memories, system of language, speech, system of oriented attention, etc (Koukolík, 2003).

Human memory has been supposed as a unique object for a long time. Nowadays it is believed that no single brain center stores memory. On the contrary, human memory is supposed as a set of cooperating (and sometimes competing) memory modules bound to different parts of human brain (Koukolík, 2003). Thus, each part of human brain contributes differently to permanent memory storage.

Cognitive science differentiates between semantic, episodic, procedural, emotional etc. memories, which "define"different kinds of semantic information. This approach seems to be interesting for the possible computer processing. If episodic memory depends on context and autobiographical events, there is a question if there is any sense to try to model this kind of semantic information. The number of possible inputs, which influence the content of this kind of memory, corresponds to whole life experience. This experience can be supposed as unique and not transferable for an individual.

If we accept an idea that episodic memory is the essential memory part for human reasoning (it means that our knowledge of outer world is not a linguistic one but it is relative - based on our perception of situation and whole-life experience), then we should ask the question if it is useful to model the other types of memories (Fig. 1), which have relationship to semantics. Then only the semantic memory (context independent memory) seems to be suitable for modeling by computer tools.
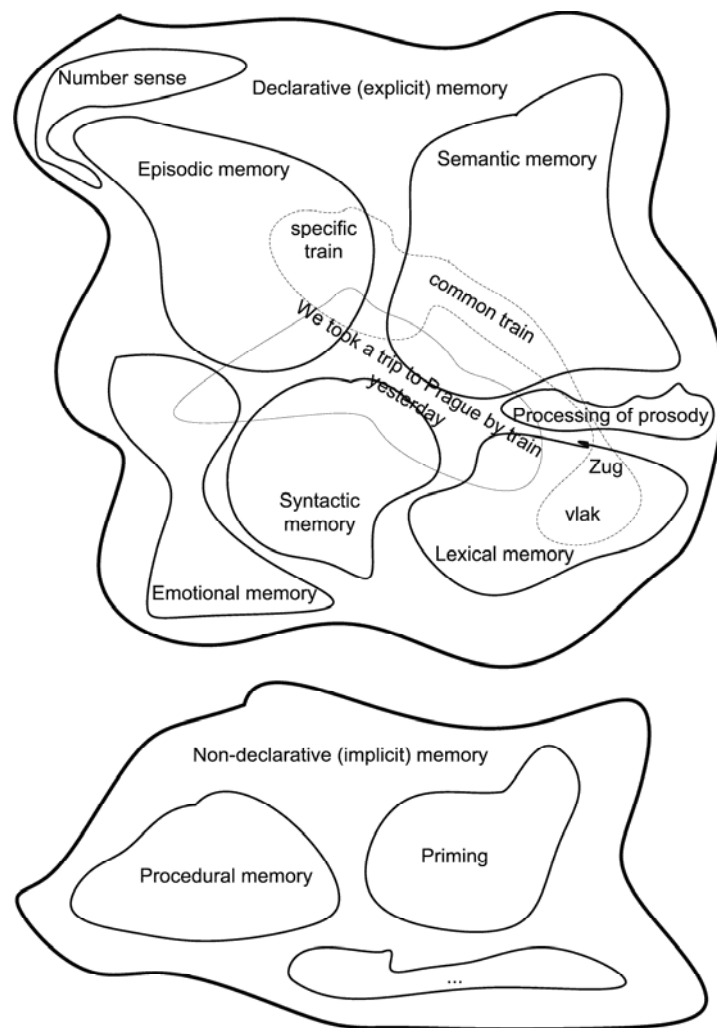
Fig. 1: Long-term memory – brain modules as independent functional systems

The part of cortex relating to semantic memory is then perceived not as a mosaic of discrete category-specific areas, but rather as a lumpy feature space representing stored information about objects features, which are shared with the members of one category. It can be said to some extent that synaptic events actively create an image of outer world proportionately to their resolving capacity. Thus, the neural representation is a highly creative activity.

2.2 Linguistics

Linguistic semantics as the scientific field investigating how the meaning is encoded in the natural language in general and how different cultures have influenced this encoding in their languages or language groups has introduced different types of meaning layers. B. Lewandovská-Tomaszcyková distinguishes two layers of meaning of simple language sign (adopted from Machová, 2001):

- Conventional layer includes prototypical, social and cultural features (corresponds to the content of semantic memory.
- Individual layer is determined with the experience of individual and also with the actual semantic features resulting from specific communication act and its context (corresponds to the content of episodic memory).

Another set of meaning layers is introduced in the work of G. Leech. A language sign has seven elementary meanings (adapted from Machová, 2001):

- Conceptual (cognitive) meaning is considered as a central base for successful human communication. It is the meaning, which is common for all the adult users of given language; it corresponds to prototypical, social and cultural features and content of semantic memory.
- Collocation meaning denotes language expressions having the same conceptual meaning but appearing only in given language collocations. Collocation meaning is hardly dependent on given language.
- Connotation meaning is an accessory meaning with respect to conceptual meaning, connotation meaning is influenced by unique life experience of individual, education, status etc. (corresponds to experience features of meaning and content of episodic memory and it usually changes in time).
- Social meaning is a kind of information about social circumstances, in which a language expression has been used.
- Emotional meaning reflects relationship of speaker to recipient and to the content of utterance (corresponds to declarative memory). Emotional meaning can be included in conceptual meaning (influences both semantic and episodic memory).
- Reflective meaning introduces a marginal component of meaning. A language sign has more than one conceptual meaning and a speaker cannot use this language sign in certain meaning without association of another meaning (a set of lexical items is smaller than a set of semantic patterns).
- Thematic-rhematic meaning is connected only with utterances. Two utterances with the same conceptual meaning differ in their communication value. This meaning is expressed with certain set of means of language in each natural language.

2.3 Artificial intelligence

The artificial intelligence also introduced several approaches how to divide semantics in layers. The human-computer dialogue systems usually cope with the sequence of the voice signal recognition (the output is the word lattice), linguistic analysis (syntactic and semantic analysis, the output is a linguistic formalism), the dialogue manager (dialogue history and the general context) and a speech synthesizer. The very important task is a building of ontology, very often in two abstract layers – the general and the domain one. The final result of this approach is then that the human-computer dialogue systems operate only in narrow domains and the final semantic representation and interpretation is reduced to selection of keywords relevant to elaborated domain.

Besides the problem with cooperation of different ontologies (and the necessity to reconstruct them in the case of change in the elaborated domain), there is an essential problem with formalisms. No formalism describes (and we believe that it cannot exist) the natural language semantics in a way that enables to process the semantics automatically. The formalisms

cannot be mapped to the natural semantics not to omit a number of expressions possible used even in a small domain and also not to produce a number of utterances without any sense.

2.4 Software engineering

If the methods of artificial intelligence are partly suitable for very specific situations, software engineering provides the methods and tools for partial semantic modelling of the world. The principles of object oriented design provide the framework how to map the semantics of real word to the structure similar to concepts of semantic memory. However, this description again takes place in different meaning layers from the description of architecture of the whole system to the description of individual components. Then the models consist of number of typed concepts, mutual associations and iterations. The principles of object oriented description provide a mechanism for interpretation of typed concepts and their features.

3. SEMANTICS AS INDIVISIBLE UNIT

The previous sections described the possible division of natural language semantics in different layers. Each field of science has introduced its own system(s) of division. The advantage of any division described above is, of course, the simplification of the problem. Moreover, we can find out that some introduced semantic layers correspond mutually across the scientific fields (e.g. the conceptual meaning and the supposed content of the semantic memory). System of layers also enables to decide that some layers are not processed at all within e.g. computerized dialogue systems. On the other side, this is a hard task to determine the exact interface between two adjacent layers (it probably does not exist) and to establish the model of communication amongst layers. Moreover, some experiments (especially in neuroscience) indicate that processing of natural language semantics in human brain is a complex process encroaching the large brain areas, in which we can find a number of bottlenecks. However, we are not able not transfer this brain model to computer to establish these natural bottleneck and system of neural communication. Then it seems that it is better to suppose the natural language semantics as an indivisible unit and process it in the smallest number of layers and steps as we are possible to do within building a computerized system.

4. PRINCIPLES AND TWO-LAYER MODEL

Research within the field of natural language processing conducted by our research group was initially based on the thought that the knowledge and natural language can be described with formalisms to a large extent (Konopík, 2004, Mouček, 2004). However, using some linguistic and AI formalisms within construction of computerized dialogue systems we were not able to get over a very limited application and a lot of routine and boring work during process of building ontologies and semantic grammars, or tagging corpora. On the other side, we had the opportunity to get first hand experience that linguistic interpretation is very dependent on individual and situation.

As the result we have decided to base our approach to the processing of natural language semantics on following principles:

- The human brain (it means neural network as an associative memory) actively creates an image of outer world.
- The linguistic interpretation of this world depends on individual and situation.

- The formalisms of traditional logic are not fully applicable, we expect only a usage of domain schemas in particular projects.
- The learning process has to be unsupervised to get rid of manual exacting work; the statistical methods will be used; supervised learning process will be used in particular project.
- In the future we expect not to use "only" the text or sound corpora but also to incorporate the other sources of knowledge as images, face expressions, biometric signals etc.

Finally our approach includes only two layers of natural language semantics processing: semantic representation layer and semantic interpretation layer. Now we work on:
- spreading of text and spoken corpora in selected domains (weather forecast, e-shopping, questions on accommodation possibilities, etc.),
- building domain schemas for selected domains,
- building software tools for supervised and unsupervised learning from corpora,
- analysis of existing approaches to natural language processing of Czech language - WEBSOM (Honkela, 1997), independent component analysis and vector state Markov models.

5. CONCLUSION

This paper described different approaches of various scientific fields to the problem of processing of natural language, especially the building of layers of semantics. We discussed the advantages and disadvantages of semantic layers and introduced our contemporary approach to natural language processing.

REFERENCES

Honkela T. (1997): Self organization maps in natural language processing. PhD. thesis. Helsinki University of Technology, Neural Networks Research Centre. P.O.Box 2200, Finland.

Koukolík F. (2003): Já (O vztahu mozku, vědomí a sebeuvědomování) (Me - Relationship between Brain, Consciousness and Self-consciousness), Charles University, Carolinum, Prague, Czech Republic.

Konopík M. (2004): Metody rozpoznávání přirozeného jazyka – návrh a implementace modulu sémantické intepretace. Diploma thesis. University of West Bohemia in Pilsen. Pilsen, Czech Republic.

Machová S. and M. Švehlová (2001): Sémantika & Pragmatická linguistika (Semantics & Pragmatic Linguistics), Charles University, Prague, Czech Republic.

Mouček, R. (2004). Sémantika v dialogových systémech. PhD thesis. University of West Bohemia in Pilsen. Pilsen, Czech Republic.