A COMBINATION OF NEURAL AND SEMANTIC NETWORKS IN NATURAL LANGUAGE PROCESSING

Andrey V. Gavrilov
Department of Computer Science, Novosibirsk State Technical University,
Nemirovich-Danchenko, 136, Novosibirsk, 630087, Russia,
Tel: +7 3832 46-04-92, fax: +7 3832 46-11-53,
Email: avg@vt.cs.nstu.ru

Abstract

The architecture of learned software for searching of semantics in text documents is proposed. In a basis of performance and the recognition of NL semantics the following fundamental principles are proposed:

1. **Orientation to a recognition of semantics with minimum usage of knowledge about syntax of the language,**
2. **Creation of hierarhies from concepts with horizontal (associative) links between nodes of these hierarhies as result of processing of text documents,**
3. **Recognition of words and collocations on maximum similar with usage of neural algorithms.**

The main algorithms of learning of software and searching of documents are considered. Also the features of learning (creation of knowledge base) of proposed software are analyzed.
Now research prototype of software with this architecture is implemented.

Keywords: natural language, artificial intelligence, neural networks, semantic networks, hybrid intelligent systems

1. Introduction

The implementation of the dialogue with the computer on the natural language (NL) is one of principal problems solved in area of computer science, called usually "«artificial Intelligence". Not without reason Turing's test, which purpose is the rating "of «quality" of an artificial intelligent system, is founded on the dialogue with the computer.
A principal problem at creation of dialogue systems using NL is the problem of recognition of semantics of sentences. Besides the solution of this problem is actual for a solution of such applied tasks, as

- Document processing (searching of semantics, referating, rubricating and so on),
- Development of retrieval servers for information retrieval in Internet with use of searching based on the natural language,
- Development of tools for transaction processing on the natural language to data bases,
- Development of testing tools in education using the open answer, in particular, in systems of distant education,
- Development of different Ask-Answer Systems, for example,

By solution of a problem of formalising and the recognition of semantics of sentences on the natural language are engaged for a long time with variable success many investigators [1-5].
However till now there is no enough complete model of performance of knowledge contained in a natural language sentence, both effective principles and algorithms of handling of the texts on NL, virtue, joining in, of knowledge engeenering methods (for example, semantic nets) and advantage of neural networks.
In paper the attempt is done to offer such model both such principles and algorithms for solving of task of search of documents.

2. Model of knowledge

In a basis of performance and the recognition of NL semantics are put the following fundamental principles:

1. Orientation to recognition of semantics with minimum usage of knowledge about syntax of the language,
2. Creation of hierarhies from concepts with horizontal (associative) links between nodes of these hierarhies as result of processing of documents,
3. Recognition of words and word collocations on maximum resembling with usage of neural algorithms.

Last two principles were formulated by author and are used by him at creation of the programming system of robots on the natural language [6].
In a figure the functional structure of system oriented to searching of documents based on their content is shown.
The editor of knowledge base – program supporting of preliminary learning of system by administrator (see steps 1 and 2 of tutoring below).
The server of documents – program for learning of system by processing of documents (see steps 3 and 4 of tutoring below).
Server of retrieves on natural language supports of searching of documents by query.
The knowledge base (KB) represents a semantic net of the frames. The frames are such named some conditionally, because ones have slots as only links to another frames-nodes of semantic net.
Conditionally in the knowledge base it is possible to select a constant component circumscribing data domain, and variable part circumscribing contents of documents. The constant part is formed as a result of of preteaching of system. The variable part is formed as a result of document processing under the control of a constant component.
The expression or group of the expressions with identical semantic, and also, word from the dictionary or document can associate with the frame. Accordingly, one of performances of the frame is its level in the net:
0 - frame coupled immediately to the word or the document (the frame-word or the frame-document),
1 - the frame, with which associates a word collocation (composite frame),
2 - frame-concept including the links on several other frames, playing the defined role in this concept,
3 - the frame-heading circumscribing concept, which is "«exposition" (link, class)) all concepts and documents coupled to this heading,
Special variety of the frames of a level 0 - frames for connection with procedures used at the analysis of NL sentences. Such frames contain words, which is processed by the special way (for example, word - "not"), or signs of punctuation - dash, colon, comma, and semicolon.
The frame consists of slots. Each slot has name and values. Value of the slot is the link to other frame, on the word in the dictionary or on the document. In the frame the following slots exist (but not all from them are used in the concrete frame):

1. Parent - link on the frame - parent or class (vertical links),
2. Owmer - list of links to frames - concepts or the composite frame, in which structure enters the given frame,
3. Obj - object participating in concept,
4. Subject - subject (or main object), participating in concept,
5. Act - operation (action) participating in concept,
6. Prop - property participating in concept,
7. Equal - list of concepts - synonyms circumscribed in the given frame (gorizontal links),
8. UnEqual - list of concepts - antonims circumscribed in the given frame,
9. Include - list of links to the frames switched on in the given concept constituent (vertical links),

Besides the frame includes the following parameters:

1. Level - level of the frame,
2. DocName - index of filename (path) of document coupled to the frame,
3. IndWord - index of a word in the dictionary coupled to the frame,
4. H - threshold of operation of the frame, as neuron,
5. Role - role of the frame in concept, which it enters or can enter (A-operation, O-object, S-subject, P-property, U-undefined or D - the operation at the analysis (is called a procedure),
6. NO - indication of inversion (refusaling) of the frame.
   A. In a system two dictionaries are used:

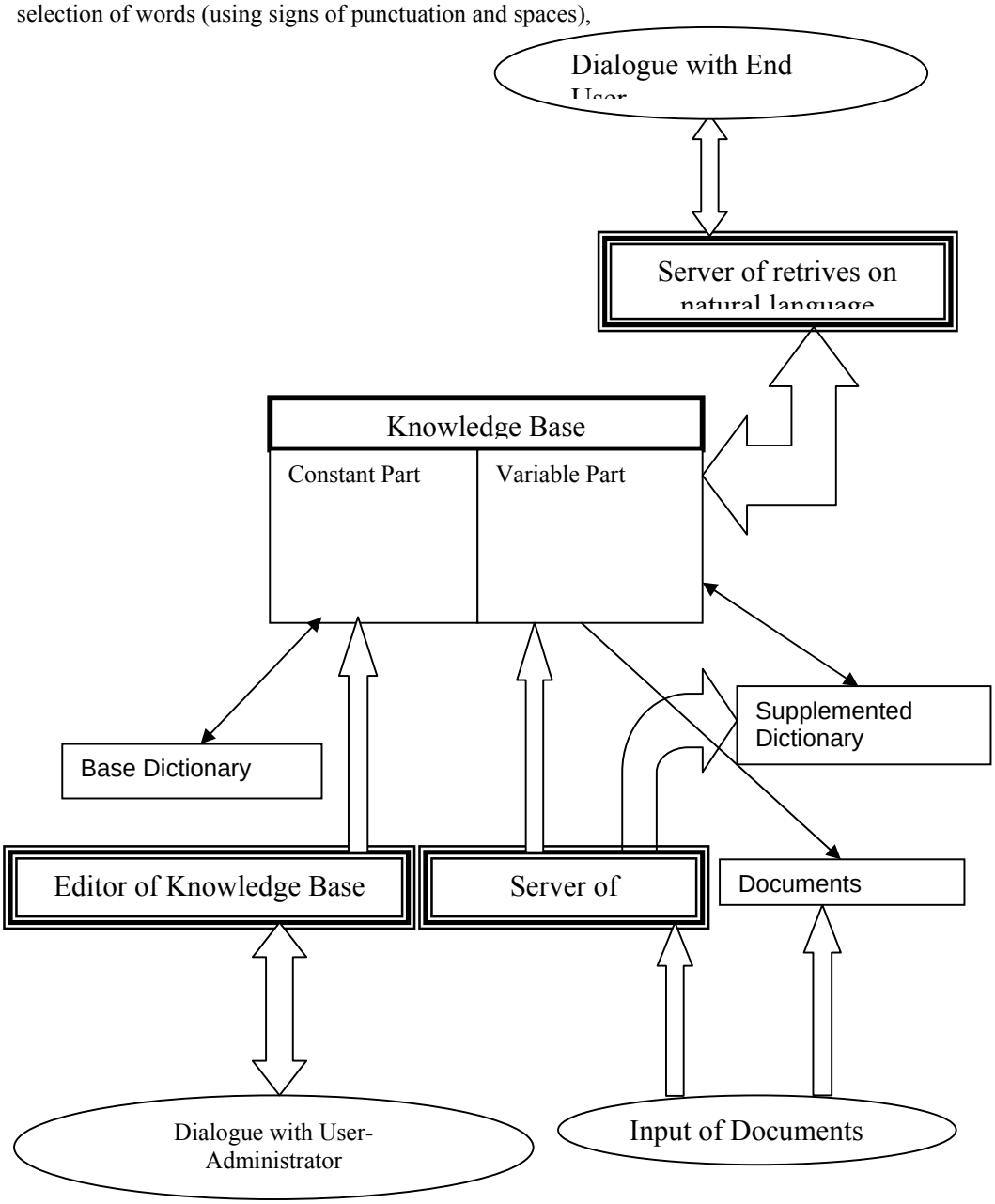
7. Basic, in which the words with their roles (essence, operation or property, in other words - noun, verb or adjective are stored,
8. The supplemented (dynamic) dictionary including a words, not recognized in the base dictionary, i.e. or absent in it (for example, names own) or present in a sentence in the too screwed aspect or in other form strongly distinguished from the form in the dictionary.

Besides some words (we shall name as their special) from these two dictionaries can be coupled (are identified) to numerals - separators (space, dash, comma). They are substituted with appropriate signs at preprocessing a sentence. It allows to filter unnecessary from the point of view of the analysis of semantic (identifying them with spaces) and to structure long word collocations and process of their analysis by replacement of such words as "«and" or "«or" on commas, "«it" or is "«equal" - on a dash, "«consists" - on a colon etc.

3. Algorithms

The processing of a sentence in a context of learning consists of the following stages:

1. selection of words (using signs of punctuation and spaces),



2. the recognition of words on maximum resembling with words in the dictionary, thus if the approaching word is not in the fundamental dictionary, then searching of this word in the supplemented dictionary, and in fail case this word adds in this dictionary.
3. the creation of the frames of a level 0, the result of this stage is object-sentence representing list of the frames,
4. replacement in this object of special words by signs-separators,
5. processing of the object-sentence by a procedure of recognition-creation of the frames.

The procedure of recognition-creation of the frames recursively processes the object-sentence, substituting word collocations between signs of punctuation on recognized in the knowledge base or created new frames (of levels 1 or 2). At the end of handling by this procedure the list of the frames in the object-sentence include only one or two frames (in the latter case in a sentence there was even one sign of punctuation "-" or colon).
The frame-concept forms on attributes in the recognized words (it probably at presence, of triples "subject" - "operation" - "object" or "object"- "operation" - "object" (in the latter case in frame - concept the first object is declared by the subject) or couples "object" - "property", "object" - "operation".
Frame - concept or composite frame form only if the appropriate frame is not retrieved in the knowledge base.
At the termination of handling in case in a sentence there is a dash, there is a creation of connections such as Equal between the frames - two parts of a sentence.
If a dash in a sentence is not present, there is a composite frame from all frames of a sentence. Last created or recognized frame is stored as a context. If the sentence starts with a dash, frame appropriate to it (recognized or created) connects to frame-context.
At creation of the composite frame or frame-concept this frame connects to frame-document by link Equal.
At handling a frame-operation to a word "not" there is an inversion of the frame-word, following it or in case "not" has met in the beginning of a sentence or at once after a dash, there is an inversion of the appropriate composite frame or frames-concept.
Frame-heading should be set before handling of the document.
The handling of a sentence in a context of retrieval processing consists of the following stages:

1. selection of words (using signs of punctuation and spaces),
2. the recognition of words on maximum resembling with words in the dictionary, thus if the approaching word is not in the fundamental dictionary, then searching of this word in the supplemented dictionary, and in fail case this word adds in this dictionary. In the latter case system ask question "what is < new word >?". The answer of the user is processed in a condition of learning.
3. the creation of the frames of a level 0, the result of this stage is object-sentence representing list of the frames,
4. the recognition of the frames of a level 1 - word collocations in the knowledge base maximum similar to recognized phrase and frames-concepts of a level 2 (here is used neural algorithm, i.e. weighed addition of signals from words, entering into the frame, or frames and matching with a threshold),
5. the searching associatively coupled by the links Equal with the recognized phrases of the frames (level 0), coupled with documents,
6. the searching of frames-documents from the retrieved frames on connections such as include, act, obj, subject, prop from above downwards, thus, if it is a lot of documents, the system produces the message with the request to reformulate sentence-request.
7. the output of the retrieved names of documents or words which are included in structure of the retrieved frames.

Frames-heading (level 3) is planned to use for abbreviation an amount of the retrieved frames. Thus those frames are sampled for the subsequent operations only which are coupled through the link Parent with appropriate by a frame-heading, or for which this link is not defined. The frame-heading can be set in the program of searching in the menu of choice of a heading or at a determination of a matching word (word collocation) in recognized sentence-request.

4. Tutoring of the program for recognition of semantics

It is recommended to train the program (to create the knowledge base) in the following sequence:
Initial tutoring to recognition of structure of sentence by the representation of sentences as "word - @symbol". After that the program before handling of a sentence at the subsequent tutoring will substitute a preset word with a preset symbol-separator. As a symbol-separator can be a space, dash, colon, comma, semicolon. A dash and colon are processed equally, Comma and semicolon too. The space means, that the given word will be eliminated from the analysis of semantics of a sentence. For example, "the - @ " means that word "the" will be ignored in during analizing of sentences, "is - @-" means that word "is" will be replaced on dash.

1. Initial tutoring. During this step the knowledge base is fullfilled by fundamental concepts from everyday practice or data domain as sentences such as "money - means of payment", "morals - rule of behaviour", "aspects of activity are set: trade, production, service" etc.

2. Base tutoring. In this step the explanatory dictionary of data domain is processed, where the concepts of any area are explained with use "-" or corresponding words.
3. Information filling. In this step the real documents are processed.

The stages 1-3 are necessary for best structurization of the knowledge base and so increasing of percent of recognized concepts at document processing at a stage 4.

5. Conclusions

A demo of the software consisting of two programs now is developed:

- for creation and debugging of the knowledge base about contents of documents,
- for documents retrieval by end user.
- The example of search, which may be used in this system, is below.

Search: «the Developers of Expert Systems».
Possible variants of contents of the documents matching to the given search are following.
Contents of the document 1: «our developments: … the expert shell ESWin»
Content of the document 2: «our company has developed an advising system based on knowledge base … »
Content of the document 3: «you can order development of the Expert System for … »
Content of the document 4: «the company is developer in software ….. Our products - ….. The toolkit for Expert Systems ESWin»
Content of the document 5: «our products …. The program of the economic analysis … based on knowledge representation by semantic nets (or rules, frames etc.) …. ».
That system could execute search, as is described above, in a condition of learning it is enough to it to meet in documents or to input in the dialogue the following sentences:

1. Expert System - advising system,
2. the methods of knowledge representation in Expert Systems: the frames, semantic nets, rules, linguistic variables.
3. the expert shell is the tool for creation and debugging of Expert Systems.

The demo is developed in Delphi 5 and is accessible for download from http://www.insycom.ru.
The software was tested on the computer with the processor AMD K6-2 350 Mh and RAM 64 Mb. In this case the knowledge base was created on the basis of the preliminary dialogue from the order 40 learning sentences, document processing (format .txt) (about 2.2 Mb).
Created KB on the carrier occupied about 8Mb (as text) and contained 87513 frames and 14715 words in the supplemented dictionary. The processing of all these documents in a condition of learning on the indicated computer proceeded within about two hours.
The developed architecture is planned to use for creation of Documents Managment System.

References

[1] P.H. Lindsay, D.A. Norman. Human information processing. - N.Y., Acad. Press, 1972.
[2] T.Vinograd. Understanding natural Language. - N.Y., Academic Press, 1972.
[3] R.C. Schank. Conceptual Information Processing. - North Holland, 1975.
[4] J. Sowa. Conceptual Structures: Information Processing in Mind and Machine. - Addison-Wesley, Ready, Massachussets, 1984.
[5] J. Yang, P. Pai, V. Honavar, I. Miller. Mobile Intelligent Agents for Document Classification and Retrieval: A Machine Learning Approach. - http://www.cs.iastate.edu/~honavar/aigroup.html, 2000.
[6] A.V. Gavrilov. A dialogue system of preparation of the programs for robots. - *Automatyka*, v. 99. - Glivice, 1988. - Pp. 173-180 (On russian).
[7] A.V.Gavrilov. An architecture of software for search of documents by natural language query. - Proc. Of Int. Worhshop KDS-2001, S.-Peterburg, 2001. - Vol. 1, Pp. 124-130 (On russian).
[8] A.V. Gavrilov. A Technology of learning for documents searching by Natural Language Query. - / The 6-th Russian-Korean International Symposium on Science and Technology. Materials. - Novosibirsk, 2002. - Vol. 3.- P.69.