Heterogeneous Semantic Networks for Text Representation in Intelligent Search Engine EXACTUS

Ivan Smirnov
Institute for Systems Analysis of RAS
Moscow, Russia
ivs@isa.ru

Ilya Tikhomirov
Institute for Systems Analysis of RAS
Moscow, Russia
matandra@isa.ru

Abstract

The paper presents an approach to text representation for search tasks. Heterogeneous semantic networks are defined and their construction from natural language is described. The success of application of semantic networks in intelligent search engine is shown.

1 Introduction

When we talk about intelligent search, we mean 'understanding' a query and finding documents relevant to the query by meaning. To understand the meaning of a query and documents we should use some semantic model which allows us to semantically match a query with documents. Since we deal with natural language documents we should use an adequate text representation model to understand them. In our opinion, heterogeneous semantic networks are a good model for text representation in search tasks.

Surely a query on natural language expresses the search request far better than simply a list of keywords does. That is why queries in the form of natural language discourse should be allowed in search engines. From this follows the necessity of semantic analysis of the query text itself and of the required documents.

Below we give a definition for the heterogeneous semantic networks, describe linguistic procedures for constructing them and their usage in search, and discuss the experimental results.

2 Heterogeneous semantic networks for text representation

In the studies devoted to the methods for representing natural language constructions, one may distinguish between the problems of lexical/lexico-semantic level representation and those of semantic/pragmatic level representation [1][2]. (By semantics one usually means the interpretation of a natural language construction in some model, e.g. in a domain model; by pragmatics - changes in the model that are initiated by this construction). The first task includes morphologic and syntactic analysis, sometimes - semantic analysis that uses the results of the lower level analysis as well as dictionary and reference information in order to construct a formalized representation of a natural language text.

Semantic level implies not on1y linguistic, but also logica1 relations between language objects to be represented [3]. Among the approaches to semantic level understanding of a text one should mention the models of 'Meaning-Text' type, preference semantics models [4], and conceptual dependence model [5]. The 'Meaning-Text' model proposes a semantic representation based on a semantic graph and description of text communicative structure.

From the standpoint of search tasks, it is important to be able to extract text elements containing searched object description and behavior laws. To such elements belong concepts,

their properties and relations between concepts. All these elements can be presented as a heterogeneous semantic network that is performed below.

2.1 Names and attributes

Name is the main unit of any description. Any language makes a distinction between individual names, general names, and meta-names. Individual names denote concrete objects of reality. General names bring into correspondence sentences and set of concepts. A general name, being the name of a set, determines the volume of a concept, collection of attributes characterizes the other side of a concept - its content.

2.2 Semantic relations

By semantic relation we shall mean a relation between concepts in the conceptual system of a domain. Parting from the typology of syntaxemes [6], we shall recognize the following kinds of semantic relations:

- Gen - generative relation, whose one component denotes a person or an object belonging to an aggregate or category denoted by the other component;
- Des - destinative relation, whose one component denotes destination of the other component;
- Dir - directive relation, where one component denotes the way (direction) of the other component;
- Ins - instrumental relation, whose one component denotes the instrument of the action denoted by the other component;
- Caus - causal relation, where one component denotes the reason of occurence of the other component some time later;
- Com - comitative relation, where one component denotes an action, object or person, accompanying the other component;
- Cor - correlative relation, whose one component expresses either the possibility of occurence of the other component or correspondence of one object to the other object or purpose;
- Neg - negative relation, where one component negates, excludes the possibility of occurence of the other component;
- Lim - limitative relation, whose one component limits the application area of the other component;
- Med - mediative relation, whose one component denotes the mode, means of the other component's action;
- Pos - possessive relation, where one component expresses the relation of possession of the other component;
- Pot - potensive relation, where one component increases the possibility of occurence of the other component some time later;
- Res - resultative relation, where one component expresses a consequence of the other component's action;
- Rep - reproductive relation, whose one component denotes the starting point for reproduction or transformation of the other component;
- Sit - situational relation, where one component denotes a situation determining the state or scope of operation of the other component;
- Fin - finitive relation, where one component means the goa1 of the other component.

In a language, semantic relations are represented by predicators, i.e. words that represent predicates. It is exactly predicators that ensure the structure of main sentences. A predicate which expresses some relation has arguments presenting components of the corresponding relation. Every argument is characterized by a semantic role in the relation. One of arguments generally has role 'subject', while the other presents specific roles of the component in the relation.

Concepts, relation and roles can be presented as a network with concepts as vertexes and relation and roles as edges.

2.3 A formal model for natural language statements representation

Let us consider an algebraic system:

$$W = (D, S, \Im, R, F), (1)$$

where,

S is some set that we shall call the set of names of objects,

R - a family of relations on S x S,

D - a universe of sets $D = \{D_1, D_m, \ldots, D_n\}$,

where each set D_i is said to be the set of attributes, and for each name $s \in S$ a subset $\Delta \in \Im$ of tuples from Cartesian product $D^k = D_1 \times D_m \ldots \times D_n$ of sets from D is related, which is referred to as the extensional or volume of the object called s.

In this case a pair $e = < s, \Delta >$ is called an event with the name s and extensional Δ. Each tuple $\delta \in \Delta$ represents a unit concept that is a concept related to a unit name.

Let F be a family of functions $\{f_1, f_2, \ldots, f_m\}$, which maps Cartesian products $D^k = D_1 \times D_m \times D_n$ of the sets from D into some of the sets D_j from D, so that for each tuple $\delta \in D^k$, a suitable function f from F relates an element $f(\delta)$ from D_j. To put it another way, the functions from F are some methods defined in an application domain that are for computing the values of certain attributes, given the values of others.

Let us consider relations from R as given not on the set of names S, but rather on events $<S_j, \Delta_j>$, so that a pair of events $<< S_1, \Delta_1 >, < S_2, \Delta_2 >>$ belongs to a relation R_m from R, if and only if the pair of corresponding names (s_1, s_2) belongs to the same relation, then (1) takes the following form:

$$W = <D, S, R, F> (2)$$

where S is the set of pairs of kind $<s_j, \Delta_j>$.

For definition of relations R_m from R see [1][2] where correspondence between the set of relations R and semantic relations of the natural language established.

3 Linguistic processing: constructing semantic networks from text

When forming a discourse, syntax deals above all with meaningful units bearing not only their individual lexical meaning, but also generalized categorical meaning in constructions of various complexity. These units called syntaxemes. Syntaxemes are minimal indivisible semantic-syntactic structures of language [6] [7]. Syntaxemes are detected taking into account: a) categorical semantics of the word; b) morphological form; c) function in the sentence (according to the constructive capabilities within a sentence there are 3 functional types of syntaxemes: free/conventional/bound).

Categorical semantics is a generalized meaning characterizing words that belong to the same categorical class (for instance, to the class of people, things, attributes for nouns).

Let us consider some examples of different syntaxemes in Russian that are similar in form, but different in meaning:

1. Мать привела сына к школе (The mother brings her son to school).
2. Ошибка врача привела к смерти (The medical error caused the death).

In (1) ‘to school’ denotes a spatial noun with the directive meaning (direction of movement), it is a free syntaxeme, since its meaning does not depend on the position in a sentence. In (2) ‘death’ denotes an attributive noun meaning logical consequence. It is a conventional syntaxeme, since its meaning is realized only in a certain complicative model in the position of a semi-predicative complicator of the model.

In a particular discourse, in a particular sentence of a query a word performs as a syntaxeme, i.e. has a certain syntactical function, in a certain grammatical form, it realizes only one of the possible meanings which this word can take in this sentence/phrase. The main task of the semantic analysis is to reveal semantic meanings of syntaxemes and relations between syntaxemes.

Following this general overview let’s now consider the linguistic analysis process in our system in more detail.

Semantic processing of the discourse is made in three stages: morphological, syntactic and semantic analysis itself. Each stage is fulfilled by a separate analyzer with its input and output data and its own settings.

3.1 Morphological analysis

At the stage of morphological analysis words and separators are recognized in the discourse. The list of all possible grammatical forms based on the word morphology is defined for each word. Word forms corresponding to the same normal dictionary form of the word and to the same part of speech and those in the same number (singular or plural) (for parts of speech that can change the number) will be classified into groups which will be further called lexemes (though they are not lexemes in the strict linguistic sense).

Obviously, a number of such lexemes can correspond to the same word. In order to reduce the number of the resulting variants of the sentence, the morphological analyzer has a filter: for every part of speech it can be defined whether or not it will be taken into account in the further analysis. The settings allow to ignore interjections and particles by default if there are variants of the word belonging to another part of speech.

In the output we get a list of sentences each of them being a list of words, and each word, in its turn, being a list of lexemes.

3.2 Syntactic analysis

The main task of the syntactic analysis is to establish syntactic dependencies between lexemes defined at the previous stage. In particular, syntactic analysis is made for extraction of minimal semantic-syntactic structures (syntaxemes).

The syntactic analysis can be done within one sentence. Compound sentences are split into simple clauses which are further processed as separate sentences. A list of variants is composed for all sentences acquired in the output of morphological analysis so that each word in each sentence variant has only one lexeme. Since the number of sentence variants is equal to the product of the number of lexemes for each word, the task of limiting the number of variants is apparent. To do so heuristics allowing to reject obviously incorrect variants are applied. Besides, a maximum allowable number of variants can be chosen in the syntactic analyzer settings.

And then the algorithm for subordinating syntactic relations discovery should be applied to each variant. As a result, lexemes are being structured into dependency trees: the lexeme at the parent node governs all child node lexemes. A minimal semantic-syntactic structure

(syntaxeme) is a tree the root of which is a noun or a preposition that governs the noun. It should be noted that proper names also belong to nouns. A noun phrase is any syntaxeme subtree into which the root noun is included.

Besides searching for revealing syntactic dependencies the syntactic analysis detects homogeneous parts. Thus at this stage two types of relations between lexemes can be detected: government and homogeneity. Every time when the program detects some syntactic relation, the relative weight of the sentence variant is increased. At the end of the sentence analysis only variants with the maximal weight are kept.

Thus the syntactic analysis input is a sentence acquired at the output of the morphological analysis. The output is a sentence in the form of a list of variants, each representing a list of dependency trees (list of syntaxemes).

3.3 Semantic analysis

The main task of the semantic analysis or, to be more exact, the semantic analysis described below (it can be deepened in the case of a properly knowledge base of the domain), is to reveal semantic meanings of syntaxemes and relations on a set of syntaxemes. In general, a semantic relation is understood as a relation of concepts in the conceptual system of the domain (see section 2). The representatives of semantic relations in lexis are predicate words, i.e. lexemes representing predicates. The verbs are predominant here, since, more often than not, they hold the central position in the semantic structure of the sentence and that influence noun phrases and sentences. The information on syntactic compatibility of every verb is recorded in special tables of relations. The tables of relations for each verb indicate types of relations between its syntaxemes (arguments).

A semantic search image consists of an ordered map of triples: <relation, arg1, arg2>, where <relation> denotes the semantic relation type, and <args> are syntaxemes, i.e. dependency trees for corresponding NPs or PPs. A semantic image can also be presented as a semantic graph with syntaxemes as vertexes and relations as edges.

To do the semantic analysis, first of all it is necessary to extract predicate words (predicators). If a verb is a predicator in the sentence, it can be detected immediately at the stage of morphological analysis. In other cases (when the predicator is a participle, a verbal noun, etc.) additional rules are applied.

When the predicate word and NPs related to it are detected, the arguments in the predicator syntaxemes structures should be filled up. The filling is made by using special linguistic dictionaries where a certain set of syntaxemes is associated with each predicator. Some rules of context and domain consideration are used to remove polysemy.

Besides, the dictionary indicates how NPs are interrelated within predicate argument structures. A set of binary relations within the set of syntaxemes is also specific for each type of predicate words and is defined a priori. The dictionary was developed by linguists and covers almost 95% of mostly frequent Russian verbs.

Let’s consider the structure of the linguistic dictionary on an example for predicate ‘love’.

Predicate= love

Meaning = subject
Syntaxeme = no preposition + subjective case
Categorical class = personal

Meaning = object
Syntaxeme = no preposition + accusative case

Categorical class = any

Meaning = causative
Syntaxeme = for + accusative case
Categorical class = attribute

Relation = CAUS
Syntaxeme1 = subject
Syntaxeme2 = causative

The collection of NPs and binary relations between them is presented in the form of a semantic graph describing the situation in the neighborhood of one predicator. This graph is a fragment of the semantic network describing the semantic image of the whole text.

Every time when a syntaxeme fills up the predicator argument or when two syntaxemes correspond to a semantic relation, the program increases the weight of the sentence variant. Hence in case of simultaneous syntactic and semantic analysis the "heaviest" variants from the point of view of both syntactic and semantic relations are chosen. That is why simultaneous analysis is not equal to sequential analysis: in the latter case variants with the greatest number of syntactic dependencies are first selected and then those ones among them are chosen in which the argument structure of predicators is filled up in a better way and more semantic relations are found. If a verb is polysemantic (i.e. there are several entries for one word in the dictionary), then all variants are considered one by one. Those variants are finally selected at the further stages of analysis, where syntaxeme meanings for a greater number of NPs of the fragment are found, and where categorial semantics attributes worked the most frequently. If there are still more than one equivalent variants again, then the variant with the maximum ratio of the number of syntaxemes found in the sentence to the total number of syntaxemes described in the given dictionary entry is to be chosen (i.e. the variant with the best (complete) verb argument structure filling).

Participial phrases and adverbial participial phrases are processed after the corresponding main clauses. The subject of the main clause becomes the subject of an adverbial participial phrase. The candidates for the subject/object of a participial phrase are the nearest NPs, whose roots agree with the participle in gender, number and case.

Syntaxemes are also searched for interrogative words like *who, what, where, why, when, how, how much, what for, at what time, etc.* A special attribute is assigned to the found meanings and when comparing the query with the document (during relevance calculation), this attribute will allow to coincide interrogative words with any NP with the same meaning in the document.

4 Using semantic networks in search

As we said above search by keywords often does not satisfy the main requirement of the user, namely the requirement of semantic relevance of the found documents to the query, even in spite of the fact that all key words of the query are present in these found documents [8].

The main idea of semantic search is semantic matching of user query with searched documents. Semantic search involves generation of semantic images of documents and queries. The semantic image in this case is presented as the semantic network so the semantic matching consists in comparison of networks vertex by vertex, role by role and relation by relation. In the result the semantic relevance is calculated that allows removing the documents obviously not semantically corresponding to the search query. Figure 1 shows examples of the semantic images of the query and the document. In this example the query includes question

word ‘who’ that is subject in relation ‘Des’ with destination ‘Brussels’. The document contains subject ‘President of Russian Federation’ and destination ‘Brussels’ which are connected with ‘Des’, so the document is semantically relevant to the query and in essence is the answer to the question.

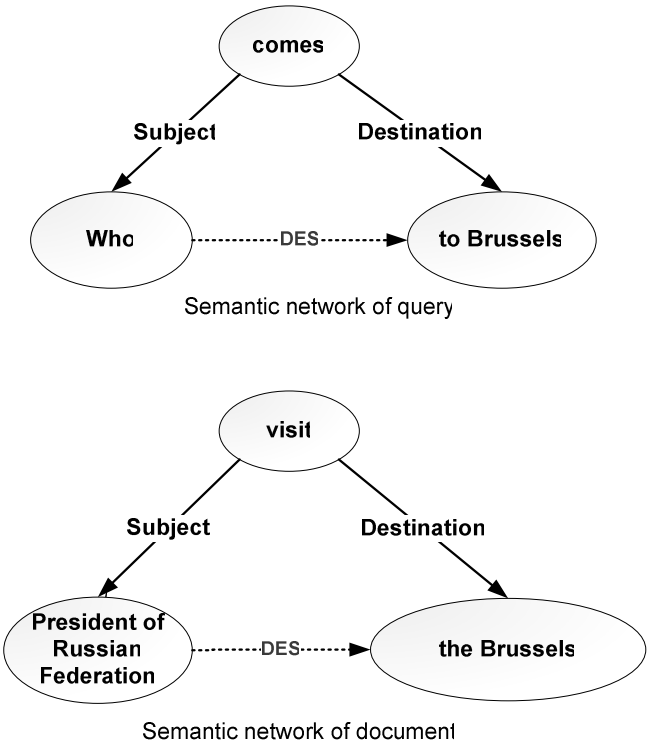

Figure 1: Semantic images of query and document.

Semantic analysis essentially enhances the precision and recall of the search and decreases the number of irrelevant documents returned in the result of the search.

5 Experimental results

The experimental search algorithm integrating traditional statistical approaches to search and linguistic processing based on principals performed above was implemented in the search engine EXACTUS and tested in ROMIP seminar in 2008 [9], [10]. In many respects ROMIP seminars are similar to other world information retrieval events such as TREC [11], CLEF, NTCIR, etc. Similar to TREC ROMIP has cycle nature and is overseen by a program committee consisting of representatives from academia and industry.

Given a collection of documents and tasks participants run their own system on the data and submit results to the organizing committee. Collected results are independently judged and the cycle ends with a workshop for discussing results and sharing experience.

Widely known metrics are used for evaluation in ROMIP:

- Precision;
- Precision at level 5 and 10;
- Average precision;
- Recall;
- 11-point TREC precision-recall graph;
- Bpref.

In 2008 EXACTUS was tested on collection of Belarusian websites and showed the best results for all metrics in OR-evaluation and for almost all metrics in AND-evaluation. Figure

2 shows TREC 11-point precision/recall graph for search algorithms EXACTUS and other participants. It is seen that EXACTUS algorithm applying semantic processing gives the highest precision/recall values. 37% of evaluated queries contain semantic information (i.e. semantic meanings and relations for words) so we can state that the semantic analysis makes important contribution to the search results.

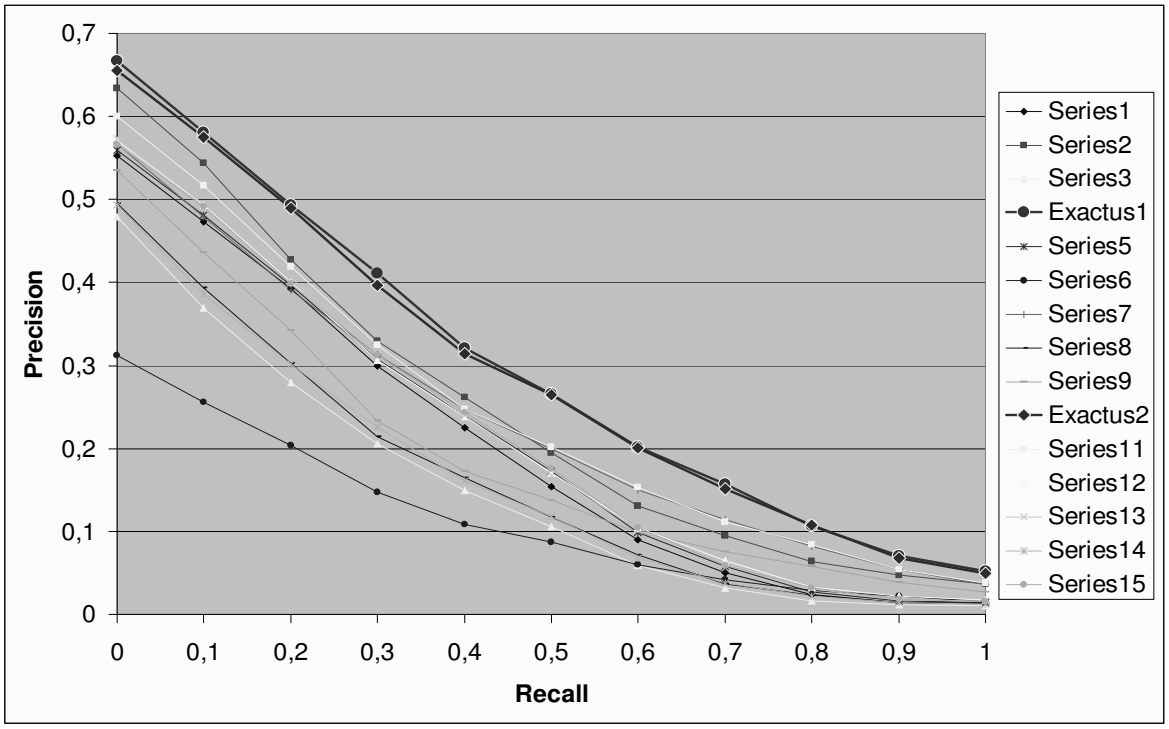

Figure 2: TREC 11-point precision/recall graph for semantic search algorithm EXACTUS

6 Conclusion and future work

We presented an approach to semantic search elaborated within an intelligent search system EXACTUS. The experiments show the advantage of using linguistic methods together with statistical methods of search for improvement of search quality.

Subjects of thorough research are methods of revealing implicit relations as well as anaphora resolution in case of multiple variants requiring semantic filtering of syntaxemes.

At present the English version of the system is being developed and the possibility of multilingual search and translation within the system is being investigated. The Russian prototype of the intelligent search engine is available at www.exactus.ru.

References

[1] G. Osipov. Semantic Types of Natural Language Statements. A Method of Representation. //10th IEEE International Symposium on Intelligent Control 1995, Monterey, California, USA, Aug. 1995.

[2] G. Osipov. Methods for Extracting Semantic Types of Natural Language Statements from Texts. //10th IEEE International Symposium on Intelligent Control 1995, Monterey, California, USA, Aug. 1995.

[3] I.A.Meltchuk, Experience in the theory of linguistic models "Meaning-Text". Moscow: Nauka, 1974. (in Russian)
[4] Y.Wilks, Preference semantics, Keenan (ed.) "Fonnal semantics of natura1 language", Cambridge, England: Cambridge University Press, 1975.
[5] R.Schank, Conceptual infonnation processing. Amsterdam: North-Holland, 1975.
[6] G. Zolotova, N. Onipenko, M. Sidorova. Communicative grammar of Russian language. Moscow, 2004. (in Russian)
[7] G.A.Zolotova. Syntactic dictionary: Repertory of elementary units of Russian Syntax. Moscow: Nauka, 1988. (in Russian)
[8] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov. Application of Linguistic Knowledge to Search Precision Improvement. // Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. - P. 17-2 - 17-5.
[9] Boris Dobrov, Igor Kuralenok, Igor Nekrestyanov, Ilya Segalovich Russian Information Retrieval Evaluation Seminar, LREC'04, february 2004 // http://www.romip.ru/docs/lrec2004-final.pdf
[10] ROMIP: Russian Information Retrieval Evaluation Seminar. http://www.romip.ru/en/index.html
[11] TREC: Text Retrieval Conference. http://trec.nist.gov/