Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts

O. Zolotarev¹, M. Charnine², and A. Matskevich²

¹Russian New University, Moscow, Russia
²Institute for Informatics Problems, Russian Academy of Sciences, Moscow

Abstract - This article discusses methods of constructing a formalized structure of a subject domain based on analysis of natural language texts, including discovering objects, their properties and related actions, followed by discovering business processes specific to the subject domain and the formation of thesaurus and business processes of the subject domain. At the same time the thesaurus can be changed based on the results of text analysis. Implementation of this approach involves automatic identification of objects mentioned in different documents, determining their properties and relationships, as well as the construction of a formalized structure of subject domain by automatic extraction information from natural language texts. The article describes methods of primary text classification in association with their particular subject domain, allowing substantially reduce the number of irrelevant documents for consideration.

Keywords: Business models, semantic networks, fragments of knowledge, objects, processes, thesaurus, Big Data

1 Introduction

This article describes the extraction of objects and business processes from the natural language texts. To develop methods for constructing models of business processes in the field of business prose assumed a detailed study of the subject domain, researching of modeling standards used as in the implementation projects of enterprise resources planning (ERP) systems as well as in projects of improving the efficiency of enterprises. During the investigation of the company dynamic and static domain models are build. This paper describes the methods of selection of objects and processes from natural language texts based on the analysis of a large volume of documents, so-called Big Data. This approach is based on development and maintenance of thesaurus of objects and processes.

One of the most labor-intensive and time-consuming phases during the implementation of enterprise resources planning systems is the initial inspection of the enterprise. Data collection methods are quite diverse. The phase of analyzing documents and subsequent construction of formalized structures of business processes are very time-consuming. The most characteristic/specific properties of the subject domain are extracted from the text documents. Automated processing of texts allows substantially reduce the costs of labor resources during the discovering information from texts as well as in formation of generalized knowledge structures inherent to the specific subject domain. Currently, various methods of natural texts processing are widely used for formalization of domain knowledge and for creation of business process models.

2 Modern investigations in the field

Syntactical analysis of the business model often reveals unused, useless functions and relationships. A more careful analysis reveals redundant and inefficient functions. One indication of the inefficiently used connections or functions is the absence of feedback loops. There is research project by Chapparro et al. (2012), which offers a selection of structured templates for extracting business rules. Business knowledge are extracted from a variety of sources - databases, structured documents, definitions, etc.


Kalsing (2010) and Nascimento (2012) propose a technique of extraction of business processes from legacy information systems by identifying business rules.

In the Rocket AeroText system the information is extracted from natural language texts based on analysis of large volumes of texts in order to discover business processes. Currently, the model of semantic vector space based on distributive hypothesis [7, 8, 4] is widely used for the analysis and processing of large volumes of natural language texts including for the automated construction of thesauri and classification systems.

Sebastyan Pado from University of Saarland (Germany), and Mirella Lapata (UK) consider construction of semantic spaces based on traditional vector models with regard to syntactic relations. Semantic properties of the words are presented in the frequency matrix, each row of
the matrix corresponds to a unique target word and each column corresponds to a linguistic context. Semantic information is extracted from a large volume of texts based on context analysis of the word. The word is considered as a point in a multi-dimensional semantic space. Semantic similarity between words is calculated based on the proximity between points of semantic space using metrics. Semantic similarity analysis is performed based on statistical methods with the calculation of the frequency of close points in the text of the semantic space. The authors use different metrics for similarity analysis: Euclidean, Jaccard's, Kullback-Leibler and others. The study concludes that context of a word is very important for discovering of lexical and semantic relations of the word.

3 Creating a model of enterprise management using analysis of domain texts

Currently there is no single definition of a subject domain. A subject domain is considered primarily from the point of view of solved tasks. Creation of a mathematical model for a subject domain with a large number of parameters is a very difficult task. The domain model should adequately reflect all the processes taking place in it. To create the domain structure it is necessary not only detailed information about the objects used in it, but also the description of relationships between objects, as well as the incorporation of external relations and restrictions. The domain model can be described as a collection of objects of the real world with the whole set of attributes and relationships, taking into account the dynamics of changes of the domain. The domain model should adequately reflect reality.

Competitiveness of any enterprise in modern society depends on choosing the right development strategy, also depends on improving the structure of enterprise management and optimization of business processes of the enterprise. Implementation of ERP (Enterprise Resource Planning) systems today is impossible without a thorough analysis of the company. Otherwise, the money spent on the implementation can be wasted. It is known that mistakes made early in the design of information systems, cost quite expensive in the later stages of implementation. One of the most time-consuming steps in the analysis of the company is the process of separation, formalization and optimization the business processes of the enterprise.

In this situation it is important to have some software tools to extract information from the documents in order to facilitate the formation of functional and data structure of the enterprise. Information required for the formation of functional structures and data structures can be extracted from documents - reports, documents, letters, business process descriptions, instructions, records, etc. These documents can be specially prepared and formalized, or can be presented as informal descriptions in natural language.

4 Statistical and classification methods used for processing of natural language texts

The phase of the company survey during the project implementation and subsequent analysis of the data collected, their structuring and formalization requires great investment of time and resources. Huge volumes of information collected has to be processed manually by a qualified technicians and expensive consultants. The model is built. It reflects current business processes of the enterprise, the so-called model "as is". The next stage is to perform analysis of this model. This model is based on the examination of documents (reports, instructions for working with legacy systems, the provisions of the enterprise, orders and so on), results of questionnaires and interviews with employees of the enterprise, and other sources.

Model parsing reveals unused, useless functions and relationships. A more careful analysis reveals redundant and inefficient functions. One the indication of inefficiently used functions is the absence of feedback management. Model-building process "as is" is very long, with a lots of iterations.

As part of the process of extracting information from natural language texts are formation of concepts and processes thesauruses. These concepts and processes thesauruses are of two types: general - to describe the common elements that are not tied to a particular domain and domain-specific taking into account the features of objects and processes specific to a given subject region. The thesauruses are built taking into account the dominance hierarchy.

During the work on the project, it is expected to use the results of other projects led by E. Kozerenko in analytic word processing and retrieval of information objects and their relationships.

This team developed a series of linguistic processors based on the unit extended semantic networks [11], including the fastest Semantix linguistic processor with built-in tone analysis texts [3].
For optimum performance, the language processor requires a fairly complete linguistic knowledge base containing linguistic and subject knowledge, which can be represented as thesauri and dictionaries. For creating linguistic knowledge base are used the results of the core team led by M. Charnine that described below, in according with the Grant RFBR ("The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems") in the field of searching, classifying, statistical and analytical processing of natural language texts (NL-texts) of very large volumes (Big Data) [4].

M. Charnine and N. Somin in the "Conceptual text generation based on key phrases" [5] described the improved method of NL-texts semantic search from the Internet with the aim of directional extracting the encyclopedic information from texts. They used not only their own achievements such as concepts key encyclopedia Keywen but well-known search engines like Google, Yandex, Yahoo, etc. and library catalogs and electronic stores as well. The collected texts handled by the statistical analysis module for the formation of the subject domain association portrait (SDAP) containing associations between specific term of the subject domain, keywords and relevant phrase. The formation algorithm prototype version of SDAP [4] tested on the material of some subject domains including business process modeling. Relevant texts founded by the algorithm SDAP were processed with the linguistic processor to build subject dictionaries and thesauri.

SDAP formation algorithm [4] supplemented with the classification methods described below, allowing more accurately select relevant texts of the subject domain.

In the article "Identification of interests of Internet users based on associative approach" [6], the author M. Sharnin, are discussed questions of definition and keyword extraction from natural language texts to conform these words to a certain category. Each keyword is assigned a weight, which is calculated taking into account the frequency of occurrence of keywords in the texts of some category. Dedicated keywords serve as the foundation for building classifiers. Category in classifiers are represented as a tuple:

\[ \langle \text{Primary keyword}, \text{category}, \text{weight} \rangle \]

There are many different measures of the degree to which two facts co-occur. We use in our work Pointwise Mutual Information (PMI) [9, 10], presented by formula (1) as follows:

\[ \text{weight} (\text{keyword}) = \text{weight} (\text{category}) = \frac{\log(p(\text{keyword} \& \text{category})/(p(\text{keyword})p(\text{category})))}{\log(p(\text{keyword} \& \text{category})/(p(\text{keyword})p(\text{category})))} \]

Formula (1)

Here, \(p(\text{keyword}\&\text{category})\) is the probability of co-occurrence of two facts: keyword exists in the document and document belongs to category. If keyword and category-i are statistically independent, then the probability that they co-occur is given by the product \(p(\text{keyword})p(\text{category})\). If they are not independent, and they have a tendency to co-occur, then \(p(\text{keyword}\&\text{category})\) will be greater than \(p(\text{keyword})p(\text{category})\). Therefore the ratio between \(p(\text{keyword}\&\text{category})\) and \(p(\text{keyword})p(\text{category})\) is a measure of the degree of statistical dependence between keyword and category. The \(\log\) of this ratio is the amount of information that we acquire about the presence of keyword when we observe document of category. Since the equation is symmetrical, it is also the amount of information that we acquire about belonging the document to the category when we observe keyword, which explains the term mutual information.

The formula (1) allows us to calculate weight of keyword associated with a given category if we know the probability of the presence of this keyword in the documents of the category.

Using of associative links between the terms for the classification purposes [6] and associative term relations with categories, calculated by the formula (1), has led to the following results:

• the complex of texts classification methods was developed; this complex uses the unit of associative links, allowing to classify the texts do not contain terms of training samples;

• the developed software implements the proposed methods of texts classification used for the selection of relevant texts on the topic of the subject domain;

• the volume of relevant NL-texts was substantially increased with the help of texts automated partitioning on the subject domain;

• the important phrases dictionaries and domain dictionaries were extended, accumulated by processing large amounts of data from the Internet.

As a result, with the help of noted above methods and programs the domain-specific knowledge bases are formed, which include thesauri (classifiers) objects and
processes, associative portraits with the specified domain knowledge processing methods for creating of domain models, fragments of semantic networks describing the relationships between processes.

5 Building a domain model based on the “process” approach

As a part of the description of models the concept of "process" and "sub-process" are regarded as interchangeable. These concepts differ only by considering different levels of processes (process - a top level or parent sub-process - subordinate level - a descendant).

At the stage of morphological analysis of texts formalized structure of sentences is built. Objects, their properties, links between objects and actions they take part in are distinguished. At the next phase phrases are formed, statistical analysis of occurrence of terms, phrases, actions in the text is spent. At the next step, the identification of the selected items is done for the entire document; associative portrait of domain-based elements extracted from the document is constructed. As a result of analysis of a representative body of texts the degree of semantic proximity of different documents is determined. So semantically similar documents relate to a particular subject area. On a given subject domain formed document library. Based on statistical analysis of documents portraits the semantically similar processes and objects are defined. Objects and processes are represented as fragments of semantic network. Consider for example the formation of fragments of a semantic network based on the analysis of natural language text with the selection of objects and processes.

The fragments of semantic network are built on the base of the following below text. The selected part of the text is sufficient for the formation of presented fragments of semantic network describing the process of repairing the vehicle:

"The car is repaired based on the order in accordance with the regulations. Manager on order examines vehicle. Because of examination is built a list of faults. According to the list of faults mechanic replaces parts. Repaired car is transferred to the owner."

To process this text with the help of the linguistic processor are used special domain vocabularies of terms included in the description of business processes, and for each term are specified its possible role (input, management, participant/mechanism, exit). For example, let’s take a number of terms of subject dictionaries representing different elements of business processes:

- name of business processes (repair, inspection, parts replacement, manufacturing of details, holding invoice, delivery);
- input (order, the list of failures, procurement, invoice, consignment note);
- management (regulations, drawing, regulation, invoice);
- participant/mechanism (master mechanic, manufactures Turner, accountant, driver);
- output (repaired auto, fault list, item, invoice, goods).

After the processing of text with the help of the linguistic processor using a thesaurus and domain-specific dictionaries the following fragments of semantic network are constructed:

- Repair (process _) (1)
- Order (Object, Input Repair) (2)
- Repaired_car (Object, Output, Repair) (3)
- Regulation (Object, Control, Repair) (4)
- Manager (Object, Member, Repair) (5)
- Manager (Object, Member, Repair) (6)

Consider the fragment (1). Dedicated process "... repaired ..." is transformed into a verbal noun "Repair."

Zero place in this fragment of a semantic network is facing the bracket element "Repair", it is extracted from the text process. The first argument indicates the type of element - the object or process. On the second argument place in the fragment, describing the process, placed the top-level process. Empty 2nd argument place in the process means that the process is a process of top-level, and he has no parent node.

On the second argument place in the fragment describing the object (2), fixed object type. If the 2nd argument
place signified (not empty), it indicates the process that owns this object.

Process models can be spitted (decomposed) to sub-processes. The process of "Repare" is split into two sub-processes "Inspection" and "Change_Parts." As a result will be constructed the following fragments of the semantic network:

- Inspection (Process, Repair) (7)
- Change_Parts (Process, Repair) (8)
- Order (Object, Inspection) (9)
- List of faults (Object, Input, Inspection) (10)
- List of faults (Object, Input, Change_Parts) (11)
- Repaired auto (Object, Output, Change_Parts) (12)

To be simple, we consider only the objects related to the inputs and outputs of sub-processes.

We can see that sub-processes (7) - (8) are executed sequentially because the inputs and outputs of the processes coincide. Fragments (9) - (12) are constructed by analogy with the already described above.

There are also some fragments that define connections between processes of the same level (in our case - one fragment):

- Link (Inspection, Change Parts) (13)

This fragment defines a sequence of sub-processes (7) and (8). This information is not redundant, because the process of determining the sequence of operations based on inputs and outputs is quite complicated and can be ambiguous.

Formation of a knowledge base (semantic network) is based on the coincidence of the parent process input and output information - on the upper level with the input of the first sub-process and output information for the last sub-process of decomposition. As a result of the analysis of natural language texts and automatic construction of a semantic network formed thesauri of objects and processes, as well as the knowledge base for a particular domain.

6 Conclusions

This paper presents a conceptual model of a system for extracting objects and business processes from natural language texts, based on the formation of thesauri and knowledge base of objects and business processes. The processing of natural language texts that describes a certain subject domain, allows us to build sets of fragments of semantic network, which form the knowledge base of the domain. Subsequent processing of incoming documents will not only allow recognize already selected objects, but also to determine their properties, relationships to other objects of the domain. A large amount of dedicated and verified knowledge fragments is able to minimize the number of errors in the development of functional structure and data structure for a given subject domain. In addition, the analysis of constructed knowledge base gives us the opportunity to create meta-knowledge that will hold general knowledge, i.e. knowledge about the features of the structure of the domain.

The described approach significantly reduces the development costs and allows optimize both the functional structure and the data structure of the enterprise. As a result it is possible to significantly reduce the costs and duration of the preparatory phase of the development of company models.

By using this method, we can greatly speed up the analysis of the domain and creating models of business processes, that is one of the first steps to modernizing the organization, improving the manageability and efficiency of its operations. This can significantly shorten the process of optimization of business processes and further restructuring of its operations.

7 Acknowledgements

This work is supported by the Russian Foundation for Basic Research, grant #13-07-00272 “The methods for automatic creation of associative portraits of subject domains on the basis of big natural language texts for knowledge extraction systems”.

8 References


