KNOWLEDGE-BASED APPROACH TO DOCUMENT ANALYSIS

Elena Sidorova, Yury Zagorulko, Irina Kononenko

Abstract: The paper presents an approach to extraction of facts from texts of documents. This approach is based on using knowledge about the subject domain, specialized dictionary and the schemes of facts that describe fact structures taking into consideration both semantic and syntactic compatibility of elements of facts. Actually extracted facts combine into one structure the dictionary lexical objects found in the text and match them against concepts of subject domain ontology.

Keywords: text analysis, ontology, natural language, fact extraction.

ACM Classification Keywords: I.2.7 Natural Language Processing - Text analysis

Introduction

The development of information systems such as intellectual document management systems or knowledge portals is one of the most actual tasks for today. This task is often considered within the framework of creating the systematized document storehouses to simplify the search for the necessary information. Despite the importance of these questions the opportunities provided by existing information systems appear to be insufficient for the intellectual organization of activity: first, it becomes difficult (practically impossible) to find the necessary information in constantly expanding archive; second, the data are often duplicated and contradict each other.

Modern information systems should be capable to solve the whole complex of tasks concerned with the management of a stream of ingoing «crude data», namely automatic classification and automatic indexing of texts, operative and adequate document routing, data transmission, storage, archiving and content -based search.

The technology is developed to automatically analyse texts of business or scientific documents in information system operating within restricted subject domains. It should provide correct addition of new documents in information space of the system and support the content-based search in it. This technology have to support adjustment of the knowledge base of the information system both in the process of its creation and during its operation [Kononenko et al., 2005].

Knowledge and data representation

The technology of text analysis uses three components of knowledge:

- ontology that includes concepts and relations of subject domain; from the point of view of the analysis the ontology describes data to be extracted from texts and placed in the database of the system;
- dictionary (thesaurus) that contains terms that represent concepts and relations of the ontology in texts;
- information content of the system, or a database.

In the system data are presented as a set of information objects (IO) of various types that describe objects of the subject domain and, in the aggregate, form information content of the system. Each IO is an instance of some element of the ontology (concept or relation) and has the structure with the fixed set of attributes specified by the expert.

Any IO may be considered as having three different aspects - structure, content, and context. The structure is characterized by a set of own attributes and attribute values. The context specifies possible environment of IO and is defined by a set of relations with other information objects. The format of IO structure and context is defined by ontology.

For example, the context of IO can be formed by the following relations of the ontology:

Part (Publication, Collection) - the relation that connects a portion to the whole (e.g., an article and a collection of articles);

Author (Person, Document) - the relation that connects a document and a creator of the document;

Publisher (Organization, Collection) - the relation that connects the book with the organization that issues it.

Besides descriptions of objects of the subject domain, the information system also contains information objects that represent various information resources, such as publication, Internet page, diagram, map, etc. The content of such resources is described by a network of the domain objects.

The technology of the analysis is aimed at processing of text information resources. Below such IOs are named *documents*.

To provide the analysis of the document text we have to perform the following actions:

- specify concept (classes) of documents and insert them in ontology;
- define the formal structure of the text for each class of documents;
- describe the schemes of the facts setting rules of extraction of facts from the text.

Formal structure of the text

In the proposed approach documents to be analyzed are information objects that are described by a certain concept (class) of the ontology, for example, *Document* class. The text representing the contents of objects of the Document class (or any other class describing a text resource) is analyzed with the purpose of extraction of the significant information, or content. The content of the document includes a set of information objects and their relations extracted from the document text.

The formal structure of the text depending on a type or genre of the document is used In the process of document analysis.

According to [Zhigalov et al., 2002] text in the digital form has at least three levels of formal structure, i.e. physical, logic and genre levels. The first one concerns presentation of the text on page, for example, by means of tags or tables of styles. The second level concerns such elements as text, paragraph, line, sentence, etc. The third level is presented by decomposition of text into genre parts. For example, the text of the business letter [Kononenko et al., 2002] includes the following genre sections: heading (sender, addressee, resume, and address), basic section (text of the letter, comments and enclosure notice), and signature.

Below any formal text structure is named as a *segment* and described by markers. The marker is defined by the list of alternative elements where an element can be:

1) a symbol or a string;

2) lexical object identified in the process of lexical analysis;

3) segment of other type.

A segment is constructed starting from following restrictions:

- *single* - the segment should not intersect with other segments of the same type; a special case of this restriction is requirement of the absence of nesting of segments;
- *min* - segment must be minimal one in the given section of text;
- *max* - segment must be maximal one in the given section of text.

The scheme of the fact

Hierarchies of classes of concepts and semantic relations defined in ontology allow one to present structure of the proposition from a subject domain in form of a fact. A set of facts constitutes propositional content of the document.

In the proposed approach the analysis is aimed at extraction of only those facts that include objects and relations of the given subject domain. The declarative description of structure of the fact and conditions (restrictions) of its extraction are named *the scheme of the fact*.

The scheme of the fact includes a set of arguments (we use only unary and binary facts) where argument can be:

- concept of ontology;
- object or class of the dictionary;
- type of the fact;
- IO of the document whose text is being analyzed.

The scheme of the fact also includes description of restrictions which are imposed on compatibility of arguments. There are semantic and structural type restrictions.

From the point of view of the result the dynamic and static schemes are distinguished. A new object (IO or the fact) is created as a result of applying the dynamic scheme; the appearance of new object can serve as a basis for application of another scheme etc. Application of the static scheme leads to changing one of the arguments, for example, IO of the document or existing object. Generally, in the course of text analysis a set of objects or relations found in the given section of text is formed.

Let us to give examples of schemes of facts:

F1: Research-Object (monument) + Locality (Western Sahara) => creation
Object-is found-in (monument, Western Sahara)

F2: Activity (work) + Object (construction project) => creation
Function (work, construction project, Kind_of_Activity: construction)

F3: Sender (Organization) + Function.Kind_of_Activity => editing
Document (Kind_of_Activity: Function.Kind_of_Activity)

Semantic restrictions

Semantic restriction is imposed on semantic characteristics of arguments of the fact. Restriction explicitly presents a pair of compatible components, where a component is a class, or a dictionary term, or the values of attribute.

For each scheme of the fact the table of semantic combinations can be generated. The table should be filled by the expert. This table is applied for:

- narrowing the set of variants of possible combinations of text units;
- accounting for mutual influence of arguments (i.e., specification of a semantic class);
- specifying attributes of resulting object.

Below we can see a little fragment of the table of semantic combinations:

Work (class) + Construction_project (class) => Work: construction

"Development" (term) + Natural_resources (class) => Work: nature management

"Development" (term) + Document (class) => Work: document creation

Structural restrictions

Besides semantic restrictions, restrictions of other language levels, such as syntactic and genre restrictions, must be considered.

For each scheme of the fact additional conditions on its arguments should be given:

- a condition on a segment, i.e. what type of a segment the arguments should be discovered within;
- position of arguments in the text (contact position, pre- and postposition, priority of positions in case of multiple choice);
- syntactic conditions (valences of terms, prepositional phrases, etc.);
- rules of combining (coordination, projectivity, maximal connectivity).

Verification of syntactic compatibility may involve simple comparison of syntactic features of terms or construction of a local syntactic dependency tree [4].

Consider an example of scheme of the fact with structural restrictions:

Fact (a1:Work, a2:Object)

- condition on a Sentence segment:;
- check valences of terms of Work class;
- check syntactic compatibility;
- search for coordinated terms;
- conform to projectivity rule;
- give priority to the postposition of Object terms relative to Work terms.

Apply this scheme to following sentence:

“It takes about 2 months to complete the installation <1> of equipments <2> and systems of automatics <3> in view of the necessary field change <4>, carrying out of production tests <5> and preparations for shipment <6> of the 2-nd diesel power stations <7>.”

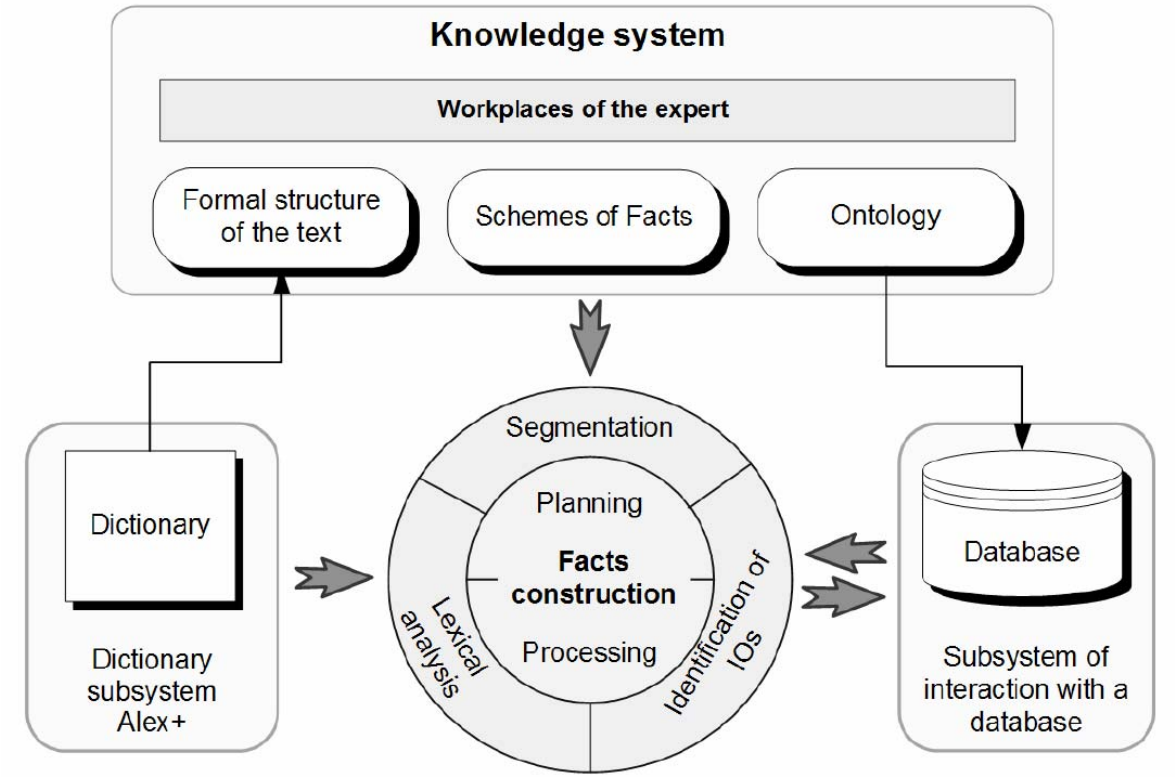
The following facts have been extracted from this sentence:

1. <1> [installation] - <2> [equipment]
2. <1> [installation] - <3> [systems of automatics y]
3. <4> [field change] - <2> [equipment]
4. <4> [field change] - <3> [systems of automatics]
5. <5> [production tests] - <2> [equipment]
6. <5> [production tests] - <3> [systems of automatics]
7. <5> [production tests] - <7> [power station]
8. <6> [shipment] - <7> [power station]

Technology of the text analysis

The system architecture (see Fig.1) includes four basic components: kernel, dictionary subsystem, editors of knowledge (ontology, schemes of facts, formal structures of the text), and a subsystem of interaction with a database.

The kernel of the system provides extraction of facts in accordance with the descriptions created by editors. The dictionary subsystem [Sidorova, 2005] ensures creation of the dictionary and realization of preliminary stage of text processing (segmentation, lexical and morphological analysis). The components realized within the project on creation of knowledge portals [Borovikova et al., 2005] are used as an editor of ontology and a module of interaction with a database.



Pic. 1. The architecture of system of the analysis.

Segmentation

There are two kinds of text segmentation - primary and genre ones.

During primary segmentation splitting linear representation of the text into ordered list of the string objects which are used for forming segments is carried out.

Genre segmentation is performed after the lexical analysis. It is based on lexical objects that mark different genre segments.

The mechanism of segmentation is realized by the Alex system [Zhigalov et al., 2002] included in the dictionary component of technology.

The lexical analysis

The lexical analysis performs extraction of lexical objects from the set of the ordered string objects obtained by the primary segmentation of the text. Lexical object is either a lexical pattern described in the Alex system, or a word/phrase represented in the dictionary.

The tasks of the given stage are following:

- application of lexical patterns;
- execution of the morphological analysis and phrase search;
- identification of genre segments.

The process results in the ordered list of objects with a following set of parameters: name (canonical form of a word or phrase, name of a pattern), position in the text, value (the main word in a synonymic group, numerical value, etc.), grammatical class (morphosyntactic information about the word form), semantic class, statistical characteristics.

Constructing facts

The mechanism of constructing facts is based on preliminary planning which is performed for each class of documents on the basis of the pre-specified schemes of facts.

Tasks of planning are the following:

1) Generation of executed rules on the basis of schemes of facts.

2) Organization of queue of rules to be executed. On this subject two aspects are taken into account:

- interdependence of schemes of facts and the order of creating objects;
- order of segments and their nesting level (the analysis is carried out starting with the smallest segment in the nesting hierarchy and proceeded up to the largest one).

3) Maintenance of correctness and convergence of process of fact construction.

During the document processing the rules are successively taken from the queue and applied. This process goes on until the queue becomes empty. For each rule data are grouped around the segments specified in a condition of a rule. Extraction of the facts is limited by frameworks of one segment.

The table of semantic combinations and syntactic rules (serving for checking of compatibility of grammatical characteristics of terms and controlling of coordination, projectivity, connectivity) are also used for fact construction.

For list of lexical objects obtained after lexical analysis the appropriate combinations are selected from the table of semantic combinations. These combinations are further considered as separate schemes of facts (however, syntactic rules are to be applied as well).

The closely adjacent objects of the same class are combined in one group. After that the contact groups are checked for compatibility (semantic and syntactic).

All the methods use the same approach to disambiguation that is based on use of weights of terms and objects. The weight depend on the following factors:

- term being a part of a phrase;
- compatibility of adjacent terms;
- term being a constituent of a fact;
- statistical characteristics, etc.

Identification of information objects

The further processing consists in forming of content of the document. For this purpose it is necessary to identify the obtained objects and provide their correct insertion into information space of the system.

The tasks of the given stage are as follows:

- Reconstruction of objects with complex structural names by means of use of “part-whole” hierarchy determined in a database;
- Reference resolution (identical objects are integrated);

- Search in a database for the objects found in the text of the document;
- Disambiguation, in case when the database includes several objects the description (content) of that corresponds to the obtained object.

The object is considered as *identified* if its class and a set of its key attributes are defined. This property allows us to distinguish the obtained object from other objects, i.e. uniqueness of objects in a database of the system is ensured.

The set of unambiguously identified objects forms a content of the document. Uniqueness of objects in the content provides its correct insertion into database of system.

Conclusion

The proposed approach is substantially based on ideas presented in [Narin'yani, 2002], in particular, we exploited idea of collaborative use of subject domain ontology and thesaurus as well as methods of semantically oriented analysis of text. In the course of practical implementation of proposed approach were also used methods and algorithms developed for experimental system for information extraction from weather forecast telegrams [Kononenko et al., 2000] and industrial intelligent document management system InDoc [Zagorulko et al., 2005].

Our immediate goals are to complete a creation of technology based on proposed approach and to apply it to solution of the laborious problem concerned with a filling of a knowledge portal with new knowledge and data [Borovikova et al., 2005].

Bibliography

[Borovikova et al., 2005] Borovikova O., Bulgakov S., Sidorova E., Zagorulko Yu. Ontology-based approach to development of adjustable knowledge internet portal for support of research activity // Bull. of NCC. Ser.: Comput. Sci. 2005. Is. 23, pp. 45-56.

[Gershenzon et al., 2005] Gershenzon., Nozhov I., Pankratov. Century System of extraction and search of structured information in big media text collections. Architectural and linguistic features. // Works of the international conference Dialogue'2005 "Computer linguistics and intellectual technologies". M.:Science, 2005, pp. 97-101. (in russian)

[Kononenko et al., 2000] Kononenko I., Kononenko S., Popov I., Zagorul'ko Yu. Information Extraction from Non-Segmented Text (on the material of weather forecast telegrams). // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2, 2000, pp.1069-1088.

[Kononenko et al., 2002] Kononenko I.S., Sidorov E.A. Business letter processing as a part of documents circulation system // Works of the international seminar Dialogue'2002 on computer linguistics and its applications. M.:Science, 2002. V.2, pp. 299-310. (in russian)

[Narin'yani, 2002] A.S. Narin'yani. TEON-2: from Thesaurus to Ontology and backwards // The international seminar Dialogue'2002 on computer linguistics and its applications. M.:Science, 2002. V.1, pp. 199-54. (in russian)

[Sidorova, 2005] Sidorova E. Technology of development of thematic dictionaries based on a combination of linguistic and statistical methods // The international conference Dialogue'2005 "Computer linguistics and intellectual technologies". M.:Science, 2005, pp.443-449. (in russian)

[Zagorulko et al., 2005] Zagorulko Yu., Kononenko I., Sidorova E. A Knowledge-based Approach to Intelligent Document Management // CSIT'2005. Ufa-Assy, Russia, 2005. V1, pp. 33-38.

[Zhigalov et al., 2002] Zhigalov Vlad, Zhigalov Dmitrij, Zhukov Alexandre, Kononenko Irina, Sokolova Elena, Toldova Svetlana. ALEX - a system for multi-purpose automatized text processing // The international seminar Dialogue'2002 on computer linguistics and its applications. M.:Science, 2002. V.2, pp.192-208. (in russian)

Authors' Information

Elena Sidorova - A.P. Ershov Institute of Informatics Systems; P.O.Box: pr. Lavrent'eva, 6, Novosibirsk, Russia, 630090; e-mail: lena@iis.nsk.su

Yury Zagorulko - A.P. Ershov Institute of Informatics Systems, P.O.Box: pr. Lavrent'eva, 6, Novosibirsk, Russia, 630090; e-mail: zagor@iis.nsk.su

Irina Kononenko - A.P. Ershov Institute of Informatics Systems; P.O.Box: pr. Lavrent'eva, 6, Novosibirsk, Russia, 630090; e-mail: irina_k@cn.ru