IMPLEMENTATION IN DECISION SUPPORT SYSTEM OF ONTOLOGICAL APPROACH TO EXTRACTION OF DATA ABOUT INFORMATION OBJECTS FROM NEWS FLOW

ALEKSANDR VALEREVICH VOLKOV
Complex Systems LLC, Skvortsova-Stepanova Street, 83, Tver, 170021, Russia
DMITRII ALEKSANDROVICH SYTNIK
Complex Systems LLC, Skvortsova-Stepanova Street, 83, Tver, 170021, Russia

ABSTRACT

The objective of this article is the study of existing approaches to extraction and structuring of data about information objects from news flow and development of the original approach to solution of that task. The article covers the approach to extraction of data about information objects based on domain otology. The developed model of information objects extracted from text is described, as well as the complex of linguistic resources applied for implementation of extraction from text of data about various types of information objects. This article gives a brief description of processing. The stage of extraction of data about information objects includes three phases: named entity recognition; inter-object relationship extraction; building complex information objects (events). For extraction of information from text, rule-based approach is used.

Keywords: *Ontology, Text Extraction, Information Extraction (IE), Rule-Based Extraction, Event Extraction, Natural Language Processing, Unstructured Data, Unformalized Data, Decision Support System, News Processing, Analysis Of News Flows, Mass Media Monitoring.*

1. INTRODUCTION

For solution of the task of gaining structured knowledge about information objects of a specific domain, it is suggested to analyze news flow for the purpose of data extraction from it. Structured knowledge about the domain's objects is especially urgent for the tasks related to subject-oriented decision support systems (hereinafter referred to as DSS). Such an example may be the system aimed at tourism sector watching which serves as illustrative material of the method described in this article.

News flow unit is an unstructured published message having the following parameters: date, time of publishing (may differ from date/time of the event) and source (mass media name). Each piece of news should contain the information about some event and its parties.

Information object is a set of data which in aggregate makes up a structured description of a real object or a phenomenon with certain level of details. For extraction of data about information objects from news flows it is suggested to apply the specific domain ontology method.

The method is based on the idea of transformation of the developed ontological model of a specific domain into a set of semantic patterns which determine the structures of the data extracted from text. Ontology of a specific domain is an accumulated database in which the relationships between various entities of that specific domain are determined which allows for texts analysis not only on lexical but on semantic level. In detail, ontological approach to information extraction from text is described in: [7], [1], [10], [16], [8].

In this article, as exemplified by tourism-related news analysis, the method of data extraction from news flows is described. In the second part of the article, information objects model is described and attributive structure of each class of information objects is given. In the third part, ontological patterns formation rules are represented and the formation procedure of the ontological portrait of a news item is given. In the fourth part, brief description of the existing approaches to data extraction from text is given and processing is described stage by stage within the described approach. The fifth part is devoted to implementation of data extraction from texts.

The suggested method was developed for the system processing texts in Russian. Nevertheless, the core ideas are true for English and the method may be applied for English texts processing.

2. MODEL OF INFORMATION OBJECTS

Information objects are described via classes. Accounting for news specifics and DSS targeting at a certain domain, it is supposed to specify two categories of classes of information objects:

- *basic* classes of information objects – classes describing information objects contained in news on any topic. Those classes are the main components of any data extraction patterns. The model of basic information objects does not depend on the domain.
- *subject-oriented* classes of information objects – classes describing specific domains. The model of subject-oriented classes is designed for each task and suggests strict modeling situation. Subject-oriented classes may be built based on basic classes via specification of more detailed items expanding the features of basic classes. Modeling subject-oriented classes may require expert's participation.

For each class of information objects, we developed special attributive structure. Values for class attributes are descriptive features which class instances may possess. Attributes may be figures, texts or instances of other classes of information objects.

The model of basic information objects describes the classes of information objects, information about which may be extracted from news texts regardless from their subjects.

From news flow it is offered to extract information about objects of the following basic classes:

a) PERSON – class describing named entity, expressing information about a certain person;
b) ORGANIZATION – class describing named entity, expressing information about a certain organization (scientific, commercial, charity, etc.);
c) GEOGRAPHIC OBJECT – class describing named entity, expressing information about a certain geographical object;
d) EVENT – class structure describing the change of information objects' state in real world. Attributive structure of Event class contains fields filled by instances of classes Person, Organization, Geographical object. An important component of Event class is temporal dimension expressed by Date field. Temporal dimension allows to state *lifetime* of the described event. The aggregate of attributive components of Event object is *event focus* – description of objects somehow interacting in news and their thematic background. Correctly set values for attributes of Event object are able to give formalized replies to the questions: *what happened* (Type of event attribute), *with who it happened* (Parties attribute), *when happened* (Time attribute). Event class is the nucleus of the basic information objects model as it aggregates the information about other information objects into uniform entity structure.

Attributive structure of basic classes is as follows:

PERSON
- Surname
- Name
- Patronymic
- Gender
- Date of birth
- Profession
- Contact information

ORGANIZATION
- Name
- Geoinformation
- Address
- Number of staff
- Contact information

GEOGRAPHICAL OBJECT
- Name
- Information from geo-objects dictionary

EVENT
- Type of event
- Time
- Place (Geographical object)
- Parties (Person, Organization)

Domain's specifics form the structure of target ontological information and cause the need for extension of the basic information objects model with extra classes describing the objects of a certain domain. Within each separate task, an expert may create a special subject-oriented model of information objects describing the classes of information units engaged in certain modeling situations determined by tasks solved by DSS.

Within the task of news analysis on tourism topic, besides the above basic classes it is useful to include some subject-oriented classes like AIR COMPANY, TOUR OPERATOR, extending ORGANIZATION class.

3. ONTOLOGY OF DOMAIN

The information objects model is developed further in domain ontology. Ontology describes the basic terms of domain and allows to set a number of ontological patterns for extraction of structured information from text. Besides information objects classes, ontology sets the relationships which may be between information objects.

For each class of information objects, a number of synonymic structures are set which, in the course of information extraction, must be identified as an instance of that class. The set of synonyms for entities is fixed by the expert manually depending on the task being solved and the domain. Information source for the list of synonyms may be Synonyms dictionary and expert's knowledge.

On the basis of information objects and relationships which may be between them, ontological patterns of extraction are created. Examples of ontological patterns extracted from tourist news messages are shown on Figure 1.

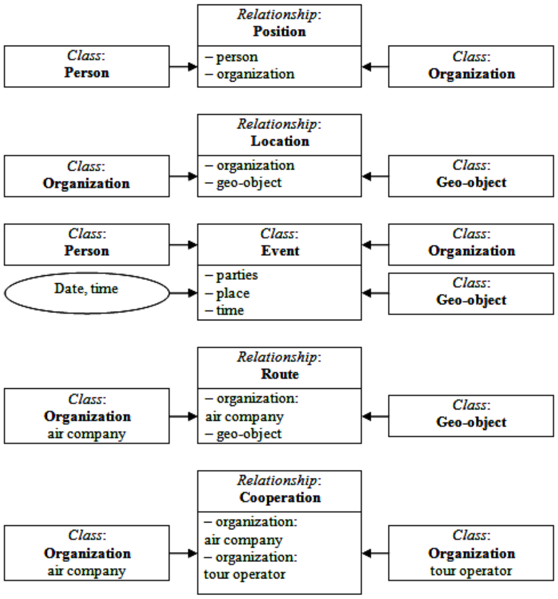

Figure 1. Ontological patterns of information extraction from news messages on tourism topic

Thus, the task of ontological patterns extraction from text is divided into two subtasks:

- task of data extraction about instances of information objects' classes. Also, this task is called the task of named entities extraction. The objective of named entities extraction task from news items is finding some or all mentions of certain persons, places, organizations and any information included in attributive structure of information objects' classes;
- task of relationships extraction between instances of information objects.

The final task of the process of data extraction about information objects based on ontological patterns from news items is the creation of ontological portrait of modeled situation.

Ontological portrait of a modeled situation is the result of uniting into a single structure of the multitudes of ontological portraits of news items. Ontological portrait of a news item is the multitude of all extracted structures of information objects compliant with ontological patterns which were managed to extract from a text represented as a table with a set of attributes strictly depending on the domain and modeled situation. The scheme of formation of ontological portrait of modeled situation is shown on Figure 2.

Ontological portrait of a modeled situation is the structured storage of actualized information interacting with DSS's model of task-solving search. Ontological portrait of a modeled situation is a structure with regularly renewed data which allows keeping them actualized.

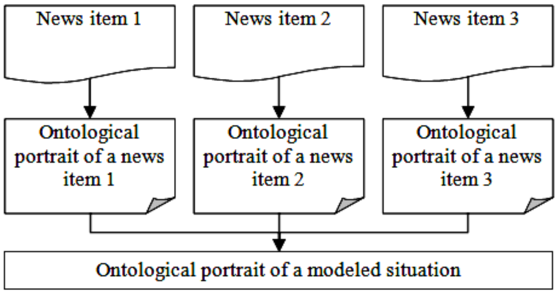
News item 1
News item 2
News item 3
Ontological portrait of a news item 1
Ontological portrait of a news item 2
Ontological portrait of a news item 3
Ontological portrait of a modeled situation

Figure 2. Scheme of formation of ontological portrait of a modeled situation

4. METHODS OF DATA EXTRACTION ABOUT INFORMATION OBJECTS

There are a few approaches to solving the task of extraction of data about information objects from unstructured texts. One of them is based on machine learning. It is based on statistical (probability) methods. Its main idea is in creation of specially laid out set of documents using which the data extraction system will be learning. The disadvantage of that approach is that it is hard to implement, as laying out documents of the multitude for learning requires a lot of time labor costs. The approach to text extraction based on machine learning is described in: [11, [12], [14], [2], [20].

There is another, rule-based approach. It is the basis of the offered method of data extraction from texts. It suggests available subsystem in application serving as the knowledge database controlling text processing. In the case described, such a subsystem is the ontology of the domain. Within that approach, rules (patterns) of information extraction from text are set manually. Rule-based approach to information extraction is studied in detail in: [3], [21], [9], [13], [19], [5], [4], [18].

The method offered includes the following stages of processing:

1.) Graphematic analysis.

Incoming text of a news item is segmented (divided) into sentences. The main indicator of sentence's end is dot punctuation character followed by space and letter symbol in the upper case. At this stage, a text is also divided into words. Normal form of each word is input in memory (using Russian dictionary of normal forms of words) and stop-words are identified. For efficient control of text resources, special technology was developed for indexing incoming documents (organization of storing text elements in computer's memory).

2.) Syntax analysis.

At this stage, each sentence is analyzed sequentially to make syntax tree for analysis. For that stage, the system uses syntax analyzer (http://www.aot.ru/). Within that process, parts of sentences and parts of speech are determined when possible as well as other grammar characteristics of words. The result of syntax analysis is making a sentence in the form of tree of dependencies. Based on dependencies between words, word combinations are detected in the text.

3.) Data extraction from texts compliant with given patterns of information extraction built on the basis of ontology.

Data extraction from text stage includes three substages:

- Extraction of named entities;
- Extraction of relationships between objects;
- Building complex information objects – events.

4.) Interpretation of data extraction results: building the ontological portrait of a modeled situation. That portrait aggregates the information obtained at the stage of data extraction from texts into uniform table structure. The set of table's attributes contains actualized information about domain's objects and serves as the data source for further implementation of intellectual data processing algorithms.

5. EXTRACTION OF DATA ABOUT INFORMATION OBJECTS

For information extraction about information objects and relationships between them from initial data – news items – it is offered to use the rule-based approach. It is based on the idea of formalization of rules regarding capital letters, numbers, punctuation marks and other attributes of text's elements which are later combined with various lists of named entities, for example, list of widespread names and surnames; geographical units; organizations; days of week; names of months, etc. Rules set the patterns in the formal language. Then, any fragments compliant with the pattern are searched in the text.

The specific of this approach is the need to write rules manually. For quality fulfillment of rule-making task for information extraction from texts of a specific domain, that domain expert's participation is useful as well as that of professional linguistic advisors.

For the analysis of tourism-related news messages, a list of priority named entities may be formed, containing names of frequently mentioned air companies, tour operators, geo-objects. Priority entities greatly increase the chances of successful identification of an object in the text. Selection of priority named entities may be done by frequency, for example, by analysis of a set of news messages accounting for extra criteria. The list of priority entities for a class is fixed by the expert and is incoming into system.

Of great significance for the result of data extraction from text is the dictionary component. The dictionary component contains the set of dictionaries and other structures for storing linguistic knowledge used for text's analysis. For news messages analysis within the approach described, at various stages of processing, certain dictionaries are used. They are dictionaries of synonyms of Russian language, frequency dictionary of Russian language, names dictionary of Russian language, word building dictionary of Russian language, subject-oriented dictionaries made with the assistance of subject's expert (dictionary of abbreviations and acronyms, dictionary of positions, dictionary of organizations, etc.).

The statistical component is also of great importance. The statistical component is responsible for collecting statistical information about objects extracted from texts and other components of the texts processed. The statistical component is an integral part of virtually any text processing system and accumulates the information about the number of words in a text, number of stop words in a text, word frequency factors and other quantity characteristics. That information is always available to the expert working with the system and allows to most fully assess the results obtained. Also, statistical parameters are actively used at the next stages of intellectual data processing, the description of which is off this article's framework.

Using the above resources together with ontological patterns of text structures extraction, information extraction quality may be greatly increased. Such comprehensive approach is more subject-oriented than the classical approach based on rules only, as auxiliary resources are adjusted in compliance with the domain.

Depending on the targets of the application developed for making patterns and applying the algorithms of data analysis, attributes of various text levels may be used. The following text levels may be specified:

- Symbol;
- Word, morpheme;
- Word combination, sentence;
- Paragraph;
- Document;
- Set of documents.

The upper text level's attributes are of great interest for making rules within the approach described: symbol, word and word combination, their attributes are actively used for writing the rules on information extraction about information objects of PERSON, ORGANIZATION, GEOGRAPHICAL OBJECT classes. Rules are written in the form of regular expressions. For creating rules, the following attributes of the above text levels are used:

- word's part of speech;
- information about lower/upper case letters in words (letter case);
- punctuation marks;
- numbers and special symbols;
- morphological characteristics (suffixes, prefixes, etc.).

5.1. Extraction of named entities

PERSON

For extraction of data about information objects of PERSON class, comprehensive extraction pattern {PERSON} is used with the following linguistic components:

- Dictionary of widespread names and parts of names;
- Set of rules describing possible objects of PERSON class in text accounting for letter case (e.g., *Ivanov Ivan Ivanovich, Ivanov Ivan, I. Ivanov,* etc.);
- Indicators of PERSON class objects. Indicator is the word which may stay either before class object or after it (e.g., *citizen, Mr., Mrs.*);
- Optionally: list of priority (widespread) named entities of Person class accounting for the task. The list of priority entities may include preinstalled set of PERSON class objects. The list of priority entities is predetermined by the task solved in DSS. That component is used in cases when the task being solved is aimed at the search and analysis of some certain entities.

ORGANIZATION

For extraction of Organization class entities, pattern {ORGANIZATION} is used with the following components:

- Set of rules describing possible variants of using punctuation marks (dots, quotation marks, etc.) and letter cases in the body of Organization class objects;
- Indicators of Organization class objects (e.g., *company, organization*). For filling the list of indicators of Organization class objects, the list of forms of companies and

organizations was used (examples of organizational forms: *OOO, ZAO,* etc.);

- Optionally: list of priority entities.

GEOGRAPHICAL OBJECT

For extraction of GEOGRAPHICAL OBJECT class entities pattern {GEO_OBJECT} is used where the information from geographical objects dictionary is input. For making geographical objects list, open dictionary of geographical objects' names was used (http://www.geonames.org/).

Geographical entities have strict hierarchy in compliance with the geographical status of an entity. The hierarchy is headed by large geographical objects: continents, oceans, seas, rivers, lakes, glaciers, deserts, etc. Further, each category is widened through smaller items like republics, cities villages, etc.

Other important component used for extraction of geographical objects is the set of acronyms (st., c., apart., sub-b., distr.), defining the rules for writing geographical entities in the text. Also, an important component is the set of rules for prepositions and particles together with geographical entities.

In the formal system's language each of the above patterns is described as {TEMPLATE_NAME} = <Class(CLASS_NAME), LC, R, PL>, where:

Class (CLASS) – the class describing attributive structure of an information object. Permissible classes: PERSON, ORGANIZATION, GEO_OBJECT;

LC – the multitude of used components of linguistic knowledge basis such as: indicators of objects auxiliary dictionaries;

R – the multitude of rules describing the patterns of extraction from the text of certain classes. The patterns are written in regular expressions format using the elements described in LC;

PL – the lists of priority named entities.

Below are the fragments of news messages on tourism topic from which the information about information objects is extracted:

– [19.05.2015 Air company (*Georgian Airlines*) since June 29, 2015 begins regular flights from Tbilisi to Yerevan.]

– [22.05.2015 Sergey Petrovich Smirnov was appointed General Director of *Rus Air* air company.]

– [06.06.2015 Air company *Orenburg Airlines* stopped cooperation with tour operator *Ideal-Tour*.]

Examples of entities which may be extracted using the information extraction patterns to those sentences:

Class: Person
Name: Sergey
Patronymic: Petrovich
Surname: Smirnov

Class: Organization
Subclass: Air company
Name: Siberian Airlines

Class: Organization
Subclass: Air company
Name: Georgian Airlines

Class: Organization
Subclass: Air company
Name: Orenburg Airlines

Class: Organization
Subclass: Tour operator
Name: Ideal-Tour

Class: Geo-object
Name: Tbilisi

Class: Geo-object
Name: Yerevan

5.2. Extraction of relationships between information objects

After the extraction of the information about named entities it becomes possible to extract relationships which may exist between objects. The task of relationship extraction between information objects is in finding all relationships between the objects mentioned in the text. Relationship means a function with one or several functional arguments, each argument being an information object. Relationship describes the type of interaction between arguments.

Relationships may be in-class and inter-class. In-class relationships are set for one class objects. In that group of relations the following types are specified:

- Type-kind;
- Synonym.

Inter-class relationships are set between various class objects. The most widespread type of inter-class relationship is represented by the relation: Part-whole (composition). Between classes, more comprehensive relationships may be set which are often hard to formalize. As a rule, such relationships are formed by experts while creating the ontology method.

For each relationship, ontology sets the rules of its extraction. The rule may be applied to text's objects within one sentence. The process of relationship extraction is built on the list of relationship indicator words. Thus, formally the structure of relationships may be described as {RELATION_NAME} = <Class(CLASS_NAME), I, S>, where:

Class (CLASS) – class describing attributive structure of an information object acting as an argument in the relationship function. Permissible classes: PERSON, ORGANIZATION, GEO_OBJECT; relationship sets the link between two or more information objects.

I – multitude of relationship indicator words used for identification of relationship in the text;

S – multitude of restrictions applied to arguments in the relationship function. Those restrictions include the rules of morphological and syntax compatibility of the relationship function's arguments.

Below, there is an example of extracted relationships between information objects:

Position: General Director = <PERSON: Sergey Petrovich Smirnov, ORGANIZATION: Siberian Airlines>

Route = < ORGANIZATION: AIR COMPANY: Siberian Airlines, GEO-OBJECT: DEPARTURE POINT: Yerevan, GEO-OBJECT: DESTINATION POINT: Tbilisi>

5.3. Building comprehensive information objects – events

EVENT is the comprehensive class describing the relationships between information objects. EVENT allows to watch the state of information objects in the real world. For extraction of EVENT class entities pattern {EVENT} is used. In the formal system's language, pattern {EVENT} is described as follows: {EVENT_NAME} = <Class(CLASS_NAME:participant), Class(CLASS_NAME:location), Relation(RELATION_NAME), Date>, where:

Class(CLASS:participant) – multitude of classes describing attributive structure of information objects being event's parties. Permissible class names: PERSON, ORGANIZATION;

Class(CLASS_NAME:location) – class describing geographical object (place) where the event occurred;

Relation(RELATION_NAME) – relationship set between objects;

Date– date of event.

Figures 3 – 5 show the schemes of extracted events from the above fragments of news messages.

Of great significance for making comprehensive information objects-events is the temporal component which sets the time of event emergence. Thus, in the example shown on Figure 3 it is advised that since May 22, 2015, Sergey Petrovich Smirnov was appointed General Director of the air company Siberian Airlines expressed in that case by Organization subclass.

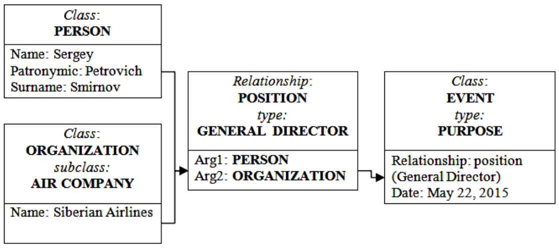

Figure 3. Scheme of extracted event Appointment to position

Figure 4 shows the scheme of extracted event – opening a new route by Georgian Airlines air company. The new route Yerevan – Tbilisi has been opened since June 29, 2015.

Figure 5 shows the scheme of extracted event – termination of cooperation. In that case the point is that the air company Orenburg Airlines stopped cooperation with tour operator Ideal-Tour. The sentence does not relate to the date of event so the date of event is considered the date of news publication, i.e. June 06, 2015.

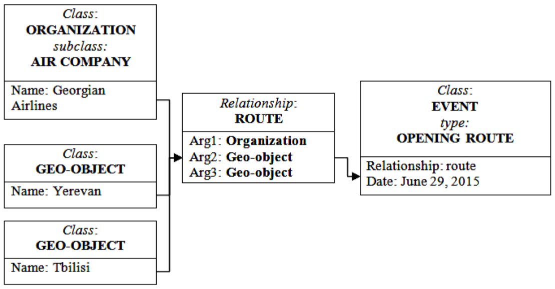

Figure 4. Scheme of extracted event Opening route

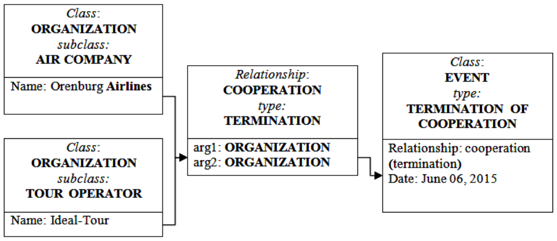

Figure 5. Scheme of extracted event Termination of cooperation

6. DISCUSSION

The main objective of this article was the development of an approach to solving tasks of gaining structured knowledge about information objects of a specific domain.

For solving the task of gaining structured knowledge about information objects, the article suggests analyzing the news messages flow. For data extraction from the flow, it is offered to use the ontological approach. That approach is based on the idea of building domain's ontology forming the space of semantic patterns (rules) extracted from the text. The main components of any ontology are the entities and links set between entities. In the approach described, information objects act as entities. The characteristics of information objects are set by the special model described herein. Also, the article describes the set of linguistic resources applied for implementation of the procedure of data extraction about various types of information objects from texts. A lot of attention is paid to the stage of data extraction about information objects which consists of three phases: extraction of named entities; extraction of relationships between objects; building comprehensive information objects (events).

Texts of news messages often contain a great part of valuable information about information objects of a specific domain. Huge volume of published news messages today makes it almost impossible to manually analyze news with the purpose to gain knowledge about domain's objects. We suggest that software implementation of the provisions described in the article will allow to greatly improve the decision quality in subject-oriented decision support systems.

Formation of structured representation of the data about information objects of some domain is the first step to creating quality-measured space of domain's information objects. Granting information objects with quantity parameters makes possible the application of quantity-based intellectual data processing methods in automatic solution search algorithms. It is important to note that information object notion is determined by semantic units, as distinct from standard text-processing approaches where terms are perceived as semiotic signs. Thus, the approach developed by us is related to the category of semantic text-processing methods.

7. CONCLUSION

The approach to data extraction from text described herein allows to form a structured data representation about specific domain's information objects. The approach is based on transformation of the developed ontological model of a domain into the set of semantic patterns which determine the text-extracted data structure.

Having structured representation of data about information objects, the system is able to apply intellectual data processing algorithms to solve DSS's tasks.

The described approach is the result of the theoretical stage of research devoted to the issues of extraction of data about information objects from unstructured sources and structuring it. Currently, software implementation of the theoretical provisions described herein is being prepared for the purpose of experiment using large scale news messages data.

Further research prospects are the development of methods to calculate numeric parameters of semantic closeness between information objects extracted from texts. Finding semantic parameters of the multitude of information objects allows for quantity comparison of ontological portraits of news messages with the portrait of a modeled situation. Thus, there will be the opportunity to quantitatively measure the actuality level of incoming news messages for the task being solved. Other variant of using numeric semantic parameters is finding the reliability level of the information extracted from texts. Solving that task will increase the quality of results of the task of data extraction from texts. The level of reliability is suggested to be determined based on frequency parameters of structures contained in news messages compliant with one ontological pattern.

8. ACKNOWLEDGMENT

This work was financially supported by the Ministry of Education and Science of the RF (Grant Agreement 14.579.21.0088, the unique identifier for applied scientific research RFMEFI57914X0088).

REFERENCES:

[1] Blythe, J., Gil, Y. (2004). Incremental Formalization of Document Annotations through Ontology-Based Paraphrasing. In *Proceedings of WWW'04 Thirteenth International World Wide Web Conference* (pp. 455-461). New York. Retrieved August 30, 2015, from http://dx.doi.org/10.1145/988672.988734.

[2] Borkar, V. R., Deshmukh, K., & Sarawagi, S. (2001). Automatic Text Segmentation for Extracting Structured Records. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 45-48). New York: ACM.

[3] Brin, S. (1998). Extracting Patterns and Relations from the World Wide Web. In *Lecture Notes in Computer Science: The World Wide Web and Databases* (pp. 172-183).

[4] Califf, M. E., & Mooney, R. J. (1999). Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)* (pp. 328-334).

[5] Ciravegna, F. (2001). Adaptive Information Extraction from Text by Rule Induction and Generalization. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI2001)* (pp. 1251-1256). Seattle.

[6] Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam. (2011). Open Information Extraction: The Second generation. In *Proceedings of IJCAI 2011* (pp. 3-10). Barcelona. Retrieved August 30, 2015, from http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-012.

[7] Gruber, T. (1995). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies, 43*(5-6), 907-928. Retrieved August 30, 2015, from http://dx.doi.org/10.1006/ijhc.1995.1081.

[8] Guarino, N. (2009). The Ontological Level: Revisiting 30 Years of Knowledge Representation. *Conceptual Modeling: Foundations and Applications Lecture Notes in Computer Science*, *5600*, 52-67. Retrieved August 30, 2015, from http://dx.doi.org/10.1007/978-3-642-02463-4_4.

[9] Guodong, Z., Jian, S., Jie, Z., & Min, Z. (2005). Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics – ACL '05* (pp. 427-434). Stroudsburg. Retrieved August 30, 2015, from http://dx.doi.org/10.3115/1219840.1219893.

[10] Hyvonen, E., Saarela, S., & Viljanen, K. (2004). Application of ontology techniques to view-based semantic search and browsing. *Proceedings of the 1st European Semantic Web symposium* (pp. 92-106). New York: Springer. http://dx.doi.org/10.1007/978-3-540-25956-5_7.

[11] Ireson, N., Ciravegna, F., Califf, M., Freitag, D., Kushmerick, N., & Lavelli, A. (2005) Evaluating Machine Learning for Information Extraction. In *Proceedings of the 22nd International Conference on Machine Learning – ICML '05* (pp. 345-352). Retrieved August 30, 2015, from http://dx.doi.org/10.1145/1102351.1102395.

[12] Knoblock, C., Lerman, K., Minton, S., & Muslea, I. (2003). Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. *Intelligent Exploration of the Web Studies in Fuzziness and Soft Computing, 111*, 275-287.

[13] Mikheev, A., Moens, M., & Grover, C. (1999). Named Entity Recognition without Gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistic* (pp. 1-8). Stroudsburg. Retrieved August 30, 2015, from http://dx.doi.org/10.3115/977035.977037.

[14] Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks: A Survey. In *The AAAI-99 Workshop on Machine Learning for Information Extraction.*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] *Natural Language Processing Modules for Syntactic and Morphological Text Analysis*. (n.d.). Retrieved August 30, 2015, from http://www.aot.ru/.

[16] Nedellec, C., & Nazarenko, C. (2006). Ontology and Information Extraction: A Necessary Symbiosis. In Buitelaar, P., Cimiano, P., & Magnini, B. (Eds.) *Ontology Learning from Text: Methods, Evaluation and Applications* (pp. 663-685), Amsterdam: IOS Press.

[17] *Online Geographical Database*. (n.d.). Retrieved August 30, 2015, from http://geonames.org/.

[18] Popescu, A.-M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the Conference on Human Language Technology and Conference on Empirical Methods in Natural Language* (pp. 339-346). Vancouver. Retrieved August 30, 2015, from http://dx.doi.org/10.3115/1220575.1220618.

[19] Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free text. *Machine Learning, 34*, 263-281. Retrieved August 30, 2015, from http://dx.doi.org/10.1023/A:1007562322031.

[20] Téllez-Valero, A., Montes-Y-Gómez, M., & Villaseñor-Pineda, L. (2005). A Machine Learning Approach to Information Extraction. *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 3406*, 539-547. Retrieved August 30, 2015, from http://dx.doi.org/10.1007/978-3-540-30586-6_58.

[21] Trampus, M., & Mladenić, D. Constructing Event Templates from Written News. (2009). *Web Intelligence and Intelligent Agent Technologies* (pp. 507–510). Milan. Retrieved August 30, 2015, from http://dx.doi.org/10.1109/WI-IAT.2009.336.