Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 145 (2018) 209–213

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society)

Building a well-formalized conceptual semantic network based on scientific and technical texts

A S Gavrilkina[a]*, O L Golitsyna[a], N V Maksimov[a]

[a]*National Research Nuclear University MEPhI, 115409, Russia*

Abstract

This paper presents a method for automatic constructing a conceptual semantic network (Task ontology) based on texts of documents. Ontology is considered as metagraph that is reflecting immanent and situational relations. Verbose constructs, which corresponds to concepts and relationships, are extracted from texts by analyzing the linguistic characteristics of words and the templates of typical constructions. Linguistic constructions of situational links are identified by relations taxonomy.

Keywords: semantic network, text processing, information retrieval, taxonomy of relations.

1. Introduction

Ontologies are presented as a result of analysis of natural language texts. They are usually associated with such areas as information retrieval, text analysis, ma-chine translation, expert systems, knowledge management systems. In recent years, many both Internet and academic publications have appeared in the field of ontology creation and research. The fundamental and practical approaches overview was given by the authors in [1, 2].

This paper presents a method for automatic constructing of a conceptual semantic network (as a Task ontology) on texts of documents. The proposed methods are based on the idea that language is an instrument of cognition and

* Corresponding author.
E-mail address: asgavrilkina@yandex.ru, nv-maks@yandex.ru

10.1016/j.procs.2018.11.042

reflection the results of mental activity. Thus the metagraph of ontology can be considered as an "assembly drawing" of the research object image (as a result of cognition).

2. Ontology as a well-formalized semantic structure

The problem of constructing an ontological description of a subject area (domain) can be formally defined as a choice of such a method of representation, which with minimum effort for perception (reading), would adequately reproduce the objects and situations of the main activity.

Ontology model, as it is proposed in [2], is considered as a set of three interrelated systems[1] (functional, conceptual and terminological in according to the semiotic model), on which the operation of comparing the elements of different systems at the level of signs is specified.

The functional system (tasks, objects, processes, properties of the subject area – sometimes named a Task ontology) represents objects and situational relations between them in the context of the target activity and is considered in direct connection with the concept used by the system, as a logical and semantic basis of ontology. In the conceptual system, the objects are stable concepts, and the set of relations is limited to generic and associative ones. The terminological system reflects the properties of the natural language at the level of signs - terms for which equivalence (synonymy) and inclusion relations are specified, as well as linguistic relations. The term is a separate word or a word-combination of a natural language.

Thus, an ontology can be formally specified as a semiotic structure defined on a set of signs that is a union of functional, conceptual and terminological systems.

From the operational point of view, graph approach can be used to define an ontology.

The functional system of ontology is represented by a weighted oriented metagraph $MG_f = \langle V_f, V_f^M, X_f \rangle$, where V_f is a set of vertices (terms), $\mathrm{V_f^M}$ is a set of metavertices, each of them can be a metagraph in turn. X_f is a set of edges (relations) defined on the set $V_f \cup V_f^M$ and $\forall x_i \in X_f : x_i = \{vb_i, ve_i, \langle tr_i, A_i \rangle\}$, where $vb_i, ve_i \in V_f \cup V_f^M$, $tr_i \in TR$ (TR is a set of relations types), $\mathrm{A_i}$ is a set of characteristic attributes of relations corresponding to edges.

Formalization of objects and links is based on the following provisions:

- the construction of representations (symbolic images) of the domain is based on the systematic nature and the conversion of objects and connections to elementary and typed entities;
- the forms and structures of the language are in some way isomorphic ("copy" but not identical) to the cognition forms. That is, the relations at the upper level correspond to the universal laws of cognition;
- text is a some "program" for the reproduction of the described object.

3. Triplets construction

Solutions for the practical construction of semantic networks (ontologies) based on text documents are presented in ISO 15926 Gellish [3] developed for the English language. However, it seems that an essential disadvantage of this approach is its focus on linguistic constructions, rather than on the model of subject area. Next, technology is focused on expert specification of facts.

The sentence of natural language, as a constructive sign, correlates with the conceivable situation described by it as its denotatum. Thus, objects of the subject area can be qualified as participants in the situation (interaction) linked by (situational) relations reflecting a particular function within the framework of an interaction specifics.

Verbose constructs, which corresponds to concepts and relationships, are extracted from texts by analyzing the linguistic characteristics of words that reflect the language properties and the templates of typical constructions. The constructs with a noun as the main word are identified as concepts. Constructions consists of verb and preposition

[1] Such model represents the knowledge generation as a self-organizing process and is based on the usage of its system property, namely the possibility of decomposition into relatively independent subsystems. That is, an object or process can be described using a set of relatively independent aspect representations. Each such description gives only partial knowledge as a whole, but complete in relation to this aspect..

are considered as relations. Pairs of concepts and relations between them form triplets, which maintains the basis for the metagraph.

4. Typification of relations

One of the features of the ontological approach that generates difficult-surmountable problems in the practical application of ontologies is that any kind of relations can be used to create ontologies. Thus, in order to be able to formally compare ontologies or to identify dependencies, it is necessary to define the classes of relations and then to identify and typify the predicate vocabulary with the help of which they can be expressed in texts in natural language.

In this direction, first of all, it is necessary to mention linguistic approaches (Y.D. Apresyan, L.M. Vasiliev, V.V. Vinogradov, Y.S. Maslov, E.V. Paducheva, Z. Wendler, W. Chaf, Ch. Fillmore, W. Cook, B. Levin). However, they are focused on the specificity of natural languages, there are no clear boundaries between classes in the classifications that are identified. Some approaches, constructed on the principle of schematism, are presented in a number of works, for example in [4-12]. There are a number of practical ontologies (DOLCE, SUMO, YATO, Gellish), FrameNet and VerbNet projects. They represent sufficiently different taxonomies of relations. All these approaches are characterized by the absence of an explicit designation and systematization of classification characteristics, which makes it difficult to use them "in pure form" and "calculate" possible dependencies through operations on ontologies.

The construction of a unified "natural" system of relations should be based on an analysis of the nature of situational connections of the objects of the subject area. In [1], a conceptual lattice of basic abstractions is proposed for the typification of relations using the following categories:

- "Concrete" / "Abstract" / "Abstract-concrete" are three basic classes reflecting the "reality / model" relations);
- "Separate" / "Singly connected" / "Multiply connected" represent the input and output relations (equal or unequal flows) as a combination of the ratios of the separate (part) and aggregate (whole).

The relations group “Abstract” of Taxonomy is associated with the artificiality (abstraightness) of the nature of at least one of linked object. The “abstract-concrete” group is conditioned by the model of interaction or correlation.

Each class of basic abstraction, in turn, is divided by "subject" (as a form of communication). There are distinguished “action-oriented” relations (analog of force, energy), “object-oriented” (action for changing the state of a target object), “result-oriented”, i.e. interaction to get a "new" object or state (analogue of an event, a fact).

Taxonomy of concepts and relationships is built on the lattice indication signs. It allows to perform operations on ontologies such as union, projection, scaling [2].

At the moment, a taxonomy of relations includes about 400 verbal constructions, as well as complementary prepositions, adverbs, nouns with a preposition and other parts of speech used in the text to express relations. Verbs and participles are reduced to the infinitive of the unreflexive form. The linguistic constructions connecting objects in triplets are distributed according to morphological templates. Modal verbs, adverbs, particles are used to characterize the modality of the relations.

5. Practical results

Pilot studies were carried out by using texts of two types of documents: design documentation (D) and standards (S). As a result, 2,342 and 5,530 triplets were obtained, in which the ratios were classified in 66% and 59% of the cases, respectively.

Table 1 provides triplets with typed relations constructed for two types of documents.

Figure 1 shows an example of a metagraph constructed for the next Russian text fragment (English translation - in brackets). *Клапаны на вертикальных сосудах следует устанавливать на верхнем днище. Клапаны не допускается использовать для регулирования давления в сосуде или группе сосудов. Изготовитель обязан поставлять клапаны с паспортом и руководством по эксплуатации. (Valves on vertical vessels should be installed on the upper bottom. Valves should not be used to pressure control in a vessel or group of vessels. The manufacturer is obliged to supply valves with a passport and operating manual).*

Figure 1 shows the class names for the relations, in square brackets there is the linguistic structure extracted from the text, the dotted lines represent hierarchical relations (which are constructed according to the lexicographical inclusion principle).

Table 1. Triplets with typed relations (S - standards, D - design documentation).

Type of document	Object 1	Object 2	Linguistic construction
class of relations "To be a means, an instrument"			
D	антиреверсное устройство (anti-reversing device)	предотвращение проворачивания вала насоса (prevention of turning the pump shaft)	служить для (serve to)
S	специальные устройства (special devices)	выливание жидких компонентов (pouring out liquid components)	применять при (apply in)
class of relations "Locativity in space"			
D	регулировочный диапазон нагрузок (adjusting load range)	диапазон 50-100 % (range of 50-100 %)	лежать в (lie in)
S	воздействие сдвоенных нормативных нагрузок (impact of dual normative loads)	расстояние 12м (distance 12 m)	устанавливать на (set at)
class of relations "Impact for object modification"			
D	оросительное теплообменное устройство (irrigation heat exchange device)	вода (water)	охлаждать (chill)
S	ороситель (sprayer)	вибрация (vibration)	подвергать (expose)
class of relations "Dependence on"			
D	Максимальное значение температуры топлива (Maximum value of fuel temperature)	величины максимального линейного теплового потока (the values of the maximum linear heat flux)	зависеть от (depend on)
S	коэффициент (coefficient)	физико-химические свойства газов (physico-chemical properties of gases)	учитывать (consider)

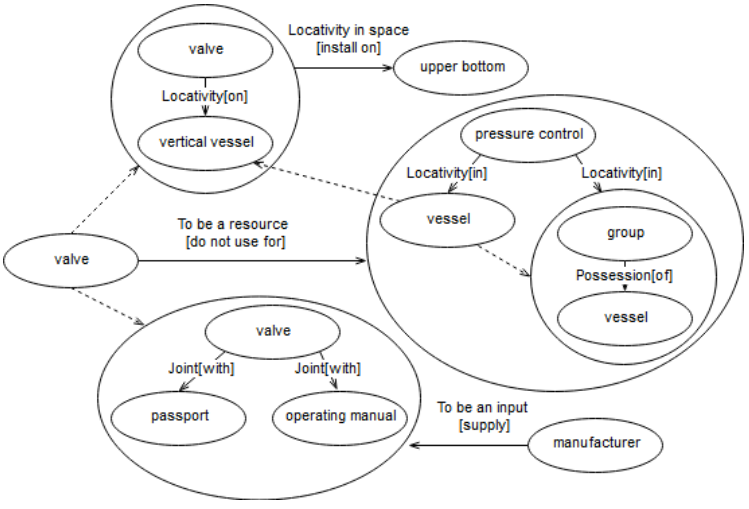

Fig. 1. Example of metagraph.

6. Conclusion

The approaches to the formalization of the semantics of texts presented in the article describing complex systems based on the structuring and typing of objects and relations of the subject area. This allows to correlate the concepts, facts and fragments of texts represented, in particular, with different verbal expressions, which is especially important in tasks of semantic search and analysis of texts. Another significant result is the inclusion of the processing of typed relations in the means of manipulating semantic images (operations on ontologies [2]) for modeling subject areas and aspects of reality. For example, to build a chain of justifications for the necessary values of parameters, to predict (more accurately, calculate) possible states and dependencies of the subject area, etc.

The dependence of the usage of typed relations and morphological templates on types of documents are studied. On a relatively small collection of documents of various types (design, standards) on "Nuclear Power Engineering" and a small array of verbal and prepositional constructions, it is shown that the constructed taxonomy covers about 70% of the relationship.

References

[1] Maksimov, N. V. (2018) "Methodological bases of ontological modeling of documentary information." *Automatic Documentation and Mathematical Linguistics* **52(2)**: 57–72.

[2] Golitsyna, O. L., Maksimov N. V., Okropishina O. V. and Strogonov V. I. (2012) "The ontological approach to the identification of information in tasks of document retrieval." *Automatic Documentation and Mathematical Linguistics* **46(3)**: 125-132.

[3] Van Renssen, A. (2005) "Gellish: A Generic Extensible Ontological Language." Delft: Delft University Press.

[4] Kitamura, Y., et al. (2002) "A functional concept ontology and its application to automatic identification of functional structures." *Advanced Engineering Informatics* **16(2)**: 145-163.

[4] Johansson, I. (2005) "Functional anatomy: A taxonomic proposal." *Acta biotheoretica* **53(3)**: 153-166.

[6] Apresyan, Yu. D. (2006) "Fundamental'naya klassifikatsiya predikatov [A fundamental classification of predicates]", in *Yazykovaya kartina mira i sistemnaya leksikografiya [Language picture of the world and systemic lexicography]*. Moscow: Yazyki slavyanskikh kul'tur.

[5] Vasil'ev, L. M. (2006) "Teoreticheskie problemy obshchei lingvistiki, slavistiki, rusistiki [Theoretical problems of General linguistics, Slavonic studies, Russian studies]." Ufa : RIO BashGU.

[8] Kitamura, Y. and Mizoguchi, R. (2004) "Ontology-based systematization of functional knowledge." *Journal of Engineering Design*. **15(4)**:. 327-351.

[9] Kitamura, Y., Koji, Y. and Mizoguchi, R. (2006) "An Ontological Model of Device Function: Industrial Deployment and Lessons Learned." *Journal of Applied Ontology (Special issue on "Formal Ontology Meets Industry")* **1(3-4)**: 237-262.

[10] Smith, B. and Grenon, P. (2004) "The Cornucopia of Formal-Ontological Relations." *Dialectica* **58(3)**: 279-296.

[11] Wood, K. L. (2002) "A Functional Basis for Engineering Design: Reconciling and Evolving Previous Efforts." *Res. Eng. Design*. **13**: 65-82.

[12] Maksimov, N. V., Okropishin A. E., Okropishina O. V. and Perederyaev I. I. (2011) "Using of Technology of Automated Formation of Conceptual Structure of Object Area of Scientific Study in Scientific Personnel Management Problems." *Vestnik RGGU. Ser. "Upravlenie"* **4(66)**: 175-185.