Development of Ontologies of Automatic Text Processing Methods Based on Ontological Design Patterns

[1] Sadirmekova Zhanna, [2] Murzakhmetov Aslanbek, [3] Ayanassova Laura, [4] Zhanar Altynbekova

[1] University of Lincoln, Lincoln, UK; Institute of Information and Computational Technologies, Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan

[2] M.Kh. Dulaty Taraz Regional University, Taraz, Kazakhstan

[3] International School of Astana, Astana, Kazakhstan

[4] Kazakh National Women's Teacher Training University, Almaty, Kazakhstan

Abstract— Currently, the volume of information on various fields of knowledge is constantly increasing. On the one hand, there is a need for automatic document processing and, on the other hand, for the emergence of new effective methods of automatic text processing in natural language and appropriate software tools and resources. Most of these methods, tools and resources are available on the Internet, but remain inaccessible to users because they are not systematized, integrated and scattered on remote pages of many websites, as well as in distributed electronic libraries and archives.

Our work is aimed at meeting the needs described above. This means the development of an intelligent information resource using modern methods of automatic text processing. The purpose of the article is to develop an ontological model of an intelligent information resource using modern methods of automatic text processing. The ontology developed in the article becomes the conceptual basis of an intelligent source of information about modern methods of automatic text processing and provides systematization of all data. Information about these methods, their integration into a single information space, convenient navigation through them and full access to them.

Index Terms— Automatic text processing, Content patterns, Domain ontology, Ontology design patterns, Methods ATP.

Introduction

Currently, ontology is recognized as the most effective means of formalizing and systematizing knowledge and data in Scientific Subject Areas (SSA) and usually understood as a subject area covering a specific scientific area or area of scientific knowledge. It is characterized in all its aspects, including research objects and subjects, used methods, the scientific activity performed and the results obtained.

The development of a domain ontology is a complex and time-consuming process. To simplify and facilitate it, various methods and approaches have been proposed [1], [2]. The following main approaches can be noted: first, to develop an ontology from scratch [3], [4]. This approach is the most time-consuming and requires the involvement of experienced specialists in the field of ontological engineering.

The second approach is to develop an ontology from ready-made blocks [5], [6], the approach involves the use of created basic ontologies and/or their fragments, which can be specialized for specific software, so it does not take much time. It allows you to involve in the process of building ontologies specialists in software for which ontologies are created.

And the third approach is the automatic construction of an ontology [7]. This approach takes the most time, but it does not allow you to build a high-quality ontology.

This paper describes an approach that implements automatic replenishment of the ontology built within the framework of the second approach. Its peculiarity is that at the first stage, knowledge engineers and experts in SSA develop and initially fill the ontology using basic ontologies and their fragments (Ontology Design Patterns or ODPs), and at the second stage - automatic replenishment of the SSA ontology with ontological entities extracted from web sites.

The first stage of the proposed approach is based on the use of ODPs [8], which are documented descriptions of proven solutions to typical problems of ontological modeling.

The use of ODPs is especially effective in the development of SSA ontologies [9]. This is due to the fact that the ODPs ontology, as a rule, contains a large number of typical fragments that are well described by ODPs. Thereby, experts in the simulated SSA who do not have the skills of ontological modeling can be involved in the development of the SSA ontology, which makes it possible to accelerate the development of the SSA ontology. In this paper, the scientific subject area is modern methods of automatic test processing (ATP).

Development An Ontology Of Modern ATP Methods

The ontology of modern ATP methods includes systematization of modern ATP methods, description of their properties and relationships between them, methods and areas of their application, publications, information resources, etc. Systematization of all information on these methods can be carried out on the following grounds: by purpose (by types of applied tasks to be solved), by areas of use.

The technology of developing intelligent information resources [10] includes the technology of developing ontologies on the basis of which these resources are built. According to the technology, these ontologies, in turn, are built on the basis of interrelated basic ontologies of scientific knowledge, scientific activity, tasks and methods, and the ontology of information resources (Fig. 1).

The ontology of scientific knowledge contains such important concepts for any scientific field of knowledge as the Object of research, Subject of research, Method of research, Scientific result and Section of science.

The ontology of scientific activity is a top-level ontology and includes basic concepts related to the organization of research activities, such as Person, Organization, Event,

Activity, Publication, used to describe participants in scientific activities, events, scientific programs and projects, various types of publications.

The ontology of Tasks and methods includes concepts describing the tasks solved within the framework of the subject area under consideration – ATP Methods, Task, Result, Terms, Application and ATP models.

In all cases, the concept of Publication plays an important role. The publication can be associated with all the concepts of software ontology. Their description is based on the basic ontology of scientific information resources, including the Information Resource class as the main class. The set of attributes and relationships of this class is based on the Dublin core standard [11].

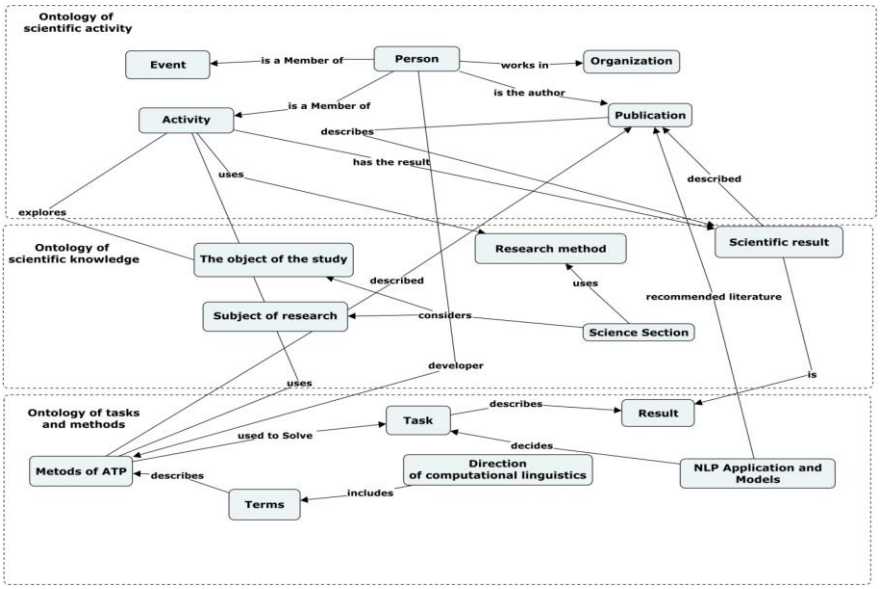

Fig. 1. Basic ontology

The core of the ontology consists of the class ATP Methods, which defines the basic properties of the ATP methods, and its subclasses used to represent the types of problems solved using methods. Such classes are Tokenization, Lemmatization, Syntactic Parsing, Semantic Analysis, Machine Learning, Information Extraction Rule-based Natural Language Processing, Statistical Natural Language Processing, Language Model, Text Classification, Part-of-SpeechTagging, Keyword Extraction, Morphological Analysis, Natural Language Generation (Fig. 2).

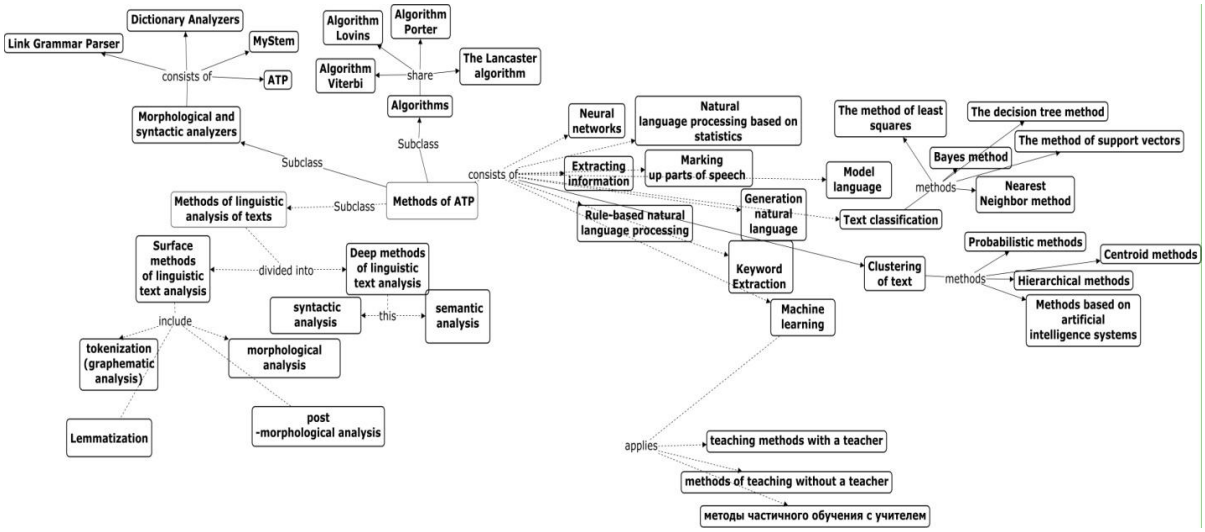

Fig. 2. Subject area ontology

The system of concepts of the subject area is formalized and the terminological core of the ontology is compiled:

– Natural Language Processing (NLP): a field of computer science and artificial intelligence that studies methods and algorithms for processing and analyzing natural language;

– Tokenization: The process of splitting text into tokens (separate units), such as words, symbols or sentences, for further processing;

– Lemmatization: the process of bringing a word to its basic form (lemma) to unify and reduce the variety of word forms;

– Syntactic Parsing: The process of analyzing the structure of sentences and determining syntactic relationships between words in order to understand their syntactic structure;

– Semantic Analysis: the process of determining semantic relations and interpreting text, including the extraction of entities, the definition of semantic roles and the construction of semantic models;

– Machine Learning: a technique that uses algorithms and models so that a computer can learn from the data and predict the results for new input data;

– Information Extraction: The process of automatically extracting structured information from unstructured text, such as extracting named entities, facts, or relationships;

– Rule-based Natural Language Processing: a text processing technique using a set of predefined rules and templates for analysis and interpretation;

– Statistical Natural Language Processing: a text analysis technique based on statistical modeling and analysis of large amounts of data;

– Language Model: a statistical model that predicts the probability of a sequence of words or symbols in a natural language;

– Text Classification: The task of defining a category or label for a given text document based on its content and context;

– Part-of-SpeechTagging: the process of determining the grammatical role of each word in a sentence (noun, verb, adjective, etc.);
– Keyword Extraction: the process of determining the most important and informative words or phrases in the text;
– Morphological Analysis: the process of analyzing and studying the internal structure of words, including their forms, endings and prefixes;
– Natural Language Generation: the process of automatic creation of text in a natural language by a computer system.

Iii. Application Of Ontological Modeling Patterns In The Development Of An Ontology Of Atp Methods

The ontology describes most fully the ATP methods implemented in the proposed IIR ATP system [12] with the use of the following patterns of ODPs: structural logical patterns, content patterns, presentation patterns and lexico-syntactic patterns.

The need to use structural logical patterns arose due to the lack of expressive means in the OWL language [13] to represent complex entities and constructions relevant to the construction of NLP ontologies, in particular, multi-place and attributed relations (binary relations with attributes), as well as the areas of acceptable values determined by the developer of the ontology. The specialization of a pattern may consist in renaming, clarifying the name and values of its properties (attributes and relationships). The specialization of patterns on the example of the structural logical pattern “Binary attributed relation" is shown in Fig. 3.

The central place in this pattern is occupied by the service class Relation with attributes, which the base classes that model the arguments of a binary relation are associated, through the links “is an Argument” and “has an Argument". At the same time, the pattern indicates that there should be one such argument at a time. The attributes of a binary attributed relationship are modeled by the properties of the Relationship class with the attributes “has an Attribute” and “has an Attribute from the Domain". In general, such a relationship may not have attributes, which is reflected in the labels of the links representing these properties. The specification of a pattern consists in substituting specific property values into it.

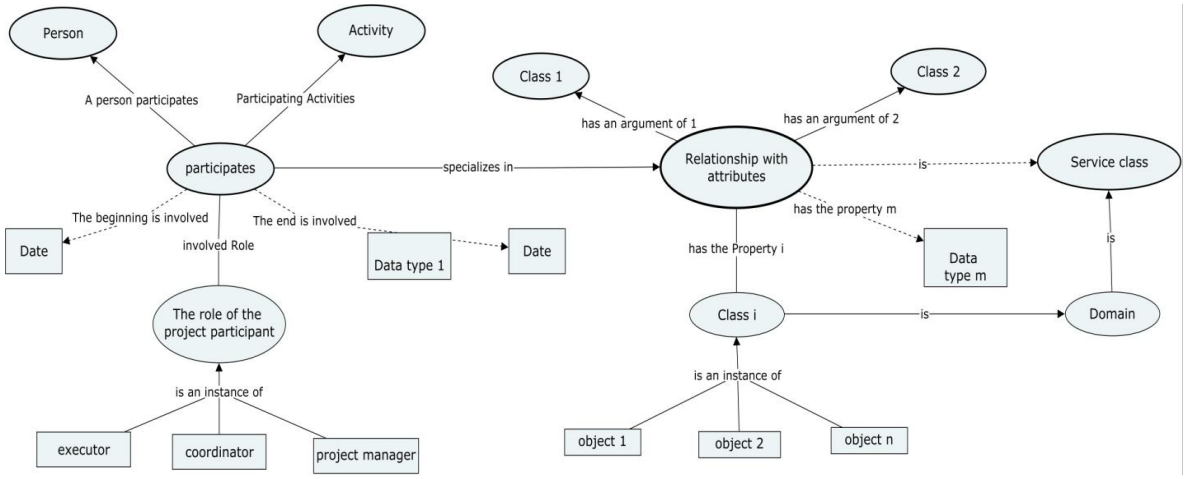

Figure 3 – Binary attributed relationship pattern

The "Range of acceptable values" pattern is designed to specify the possible values of a class property, when the entire set of such values (usually string values) is known in advance and therefore can be fixed at the development stage.

Content patterns are designed to support a uniform and consistent representation of the concepts used in NLP and their properties. Such patterns were developed for the concepts characteristic of most

SSA: The object of research, the Subject of research, Method, Task, Section of science, Scientific result, Activity, Project, Person, Organization, Publication, Information resource, etc. For each of these patterns, a set of competence testing questions is defined. With the help of these questions, the mandatory and optional compositions of the ontological elements of the pattern are identified and the requirements for them are described, which are presented in the form of axioms and restrictions.

The ontology of modern methods of automatic text processing is implemented in the Protégé 5.5.0 editor [14] (Fig. 4).

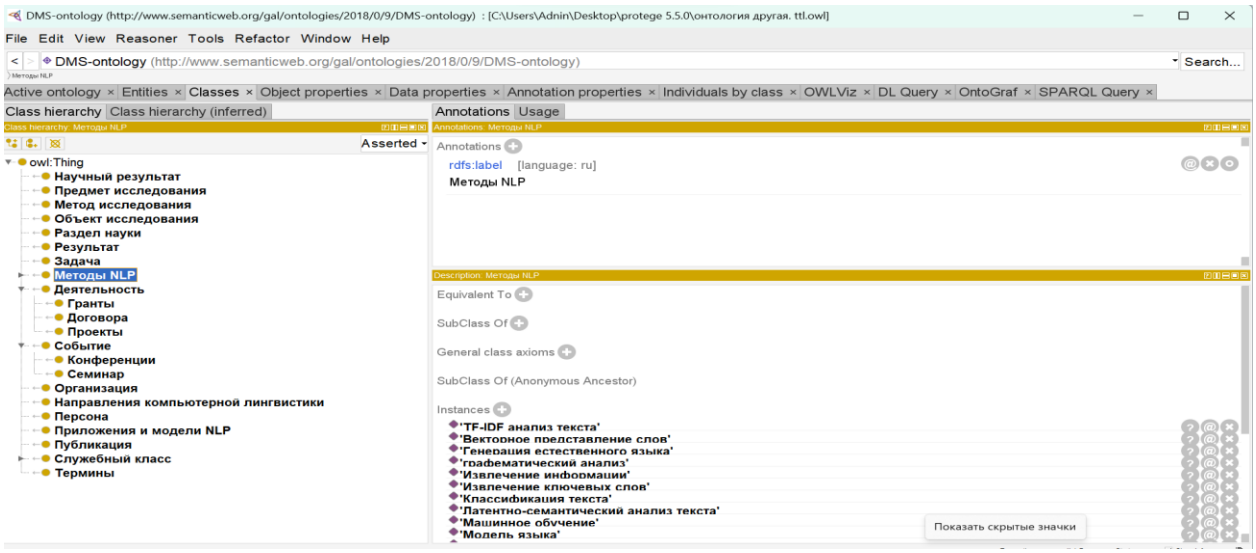

Fig. 4. Ontology editor

Extracting information about instances and relationships between them occurs using information extraction patterns (Fig. 5). The template is an XML document in which markers are specified for objects, relationships, and attributes of the ontology, signaling the location of this object, relationship, or attribute on an HTML page. Also, for each object in the template, a specialized handler can be specified, designed to extract information about objects of a certain ontology class. It is important to note that information about entities can be set in various ways. A separate template is built for each of these presentation methods.

```
<templates ontology="http://www.semanticweb.org/IIS/ontologies/INIR_Ontology#">
<class class_id=":Проект">
    <marker template="проект" element="Menu" target_element="Page" />
    <marker template="проект" element="Head" target_element="Block" />
    <attr attr_id=":Название" type= "string">
        <marker template="проект" element="Head" target_element="Head" />
        <marker template="проект" element="sentence" target_element="QuoteText" />
        <marker template="название_проекта" element="sentence" target_element="sentence" />
    </attr>
    <attr attr_id=":Номер" type= "string">
        <marker template="номер_проекта" element="sentence" target_element="sentence" />
    </attr>
    <relation relation_id=":являетсяПубликацией_Публикация_Деятельность" >
        <marker template="публикации" element="Menu" target_element="Page" />
        <marker template="публикации" element="Head" target_element="Block" />
        <class class_id = ":Публикация" engine ="PublicationsList" />
    </relation>
</class>
</templates>
```

Fig. 5. Text extraction template

CONCLUSION

This paper describes an approach that implements automatic replenishment of the ontology. At the first stage, together with Scientific Subject Areas (SSA) experts, a basic ontology was developed using Ontology Design Patterns or ODPs. The ODPs used in this approach appeared as a result of solving the problems of ontological modeling that the authors of the paper encountered in the process of developing ontologies for various scientific subject areas. The use of ontological design patterns makes it possible to ensure a uniform and consistent representation of all the entities of the SSA ontology, reduce the number of errors in ontological modeling, increase the "comprehension" of the ontology by developers and thereby enable the collective development of ontologies.

At the second stage - automatic replenishment of the SSA ontology with ontological entities extracted from web sites. A method of extracting information from web pages has been developed that combines machine learning methods and template-based methods. These templates are created based on the

ontology and take into account the structure of web documents. Multiple templates can be combined together to extract information about related objects. However, creating these templates is a rather time-consuming task, full automation of this process requires further research.

The paper was supported by a grant to finance scientific, scientific and technical projects for 2022-2024. By the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (AP14972834).

REFERENCES

[1] Sattar, A. Comparative Analysis of Methodologies for Domain Ontology Development: A Systematic Review / A. Sattar, E. Salwana, M. Surin, M.N. Ahmad, M. Ahmad, A.K. Mahmood // International Journal of Advanced Computer Science and Applications. 2020. Vol.11(5). P.99–108.

[2] Noy, N. Ontology Development 101: A Guide to Creating Your First Ontology / N. Noy, D. McGuinness // Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[3] Brusa, G. Towards ontological engineering: a process for building a domain ontology from scratch in public administration/ G. Brusa, M. L. Caliusco, and O. Chiotti // Expert Systems. 2008. Vol.25. P.484-503.

[4] De Nicola, A. A Lightweight Methodology for Rapid Ontology Engineering / A. De Nicola, M. Missikoff // Com. ACM. 2016. Vol.59. P.79–86.

[5] Gangemi, A. Ontology Design Patterns / A. Gangemi, V. Presutti // In: Staab S., Studer R. (eds) Handbook on Ontologies. IHIS. - Springer, Berlin, Heidelberg, 2009. P.221–243.

[6] Batyrkhanov A.G., Sadirmekova Zh.B., Sambetbayeva M.A. Nurgulzhanova A.N., Ismagulova Z.S., Yerimbetova A.S. Development of methods and technologies for creating intelligent scientific and educational internet resources//Bulletin of Electrical Engineering and Informatics. – 2022. – Vol. 11 (5). – P. 2968–2977. (Q3, 50%). DOI: 10.11591/eei.v11i5.3075

[7] Sadirmekova Zh.B., Tussupov J.A., Sambetbayeva M.A., Altynbekova Zh.T. Development of integrated information systems to support scientific activity // SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies, art. No. 9465964. DOI**:** 10.1109/SIST50301.2021.9465964

[8] Ontology Design Patterns (ODPs) Public Catalog, 2009. URL: http://odps.sourceforge.net

[9] Sadirmekova Zh., Tussupov J., Murzakhmetov A., Zhidekulova G., Tungatarova A., Tulenbayev M., Akhmetzhanova Sh., Altynbekova Zh., Borankulova G. Ontology engineering of automatic text processing methods // International Journal of Electrical and Computer Engineering (IJECE)– 2023. – Vol. 13 (6). – P.6620-6628. DOI: 10.11591/ijece.v13i6.pp6620-6628

[10] Zhanna Sadirmekova, Madina Sambetbayeva, Elmira Daiyrbayeva, Aigerim Yerimbetova, Zhanar Altynbekova, Aslanbek Murzakhmetov Constructing the Terminological Core of NLP Ontology //8th International Conference on Computer Science and Engineering. – Burdur-Turkey, 2023.– P. 81–85.

[11] Sadirmekova Zh.B., Zhizhimov O.L., Tussupov D.A., Sambetbayeva M.A. Requirements for information system to support scientific and educational activities // CEUR Workshop Proceedings . – Novosibirsk, Russia, 2019. – P. 44–47.

[12] Sadirmekova Zh., Sambetbayeva M., Serikbayeva S., Borankulova G., Yerimbetova A., Murzakhmetov A. Development of an intelligent information resource model based on modern Natural Language Processing methods // International Journal of Electrical and Computer Engineering (IJECE) – 2023. – Vol. 13 (5). – P. 5314-5332. DOI: 10.11591/ijece.v13i5.pp5314-5332

[13] Web page: W3C Semantic web. Available online: https://www.w3.org/2001/sw/wiki/Main_Page (accessed on 25 February 2023).

[14] Protégé. A free, open-source ontology editor and framework for building intelligent systems. Available online: http://protege.stanford.edu/ (accessed on 18 February 2023).