Knowledge Portal on Computational Linguistics: Content-Based Multilingual Access to Linguistic Information Resources

YURY ZAGORULKO, OLESYA BOROVIKOVA, GALINA ZAGORULKO
A.P. Ershov Institute of Informatics Systems
Siberian Branch of the Russian Academy of Sciences
Lavrentiev str., 6, Novosibirsk
RUSSIA
zagor@iis.nsk.su http://www.iis.nsk.su/index_e.shtml

Abstract: - The paper presents a knowledge Internet portal that provides systematization of knowledge and information resources on computational linguistics and their integration into a uniform information space, as well as the content-based multilingual access to them (information search in terms of the subject domain and knowledge-based navigation through the portal content).

Key-Words: - Knowledge Portal, Ontology, Information Resource, Content-Based Access, Knowledge-Driven Navigation, Computational Linguistics

1 Introduction

Constantly growing needs for facilities of automatic document processing and for natural language interfaces, including interfaces with speech input and output, raise the problem of efficient access not only to publications describing the methods of and approaches to text and speech processing, but also to various kinds of vocabularies, software components and algorithms which realize diverse tasks of text and speech processing. Though a great volume of knowledge and information resources relating to this area of knowledge are accumulated and presented in the Internet, access to them is rather complicated as they are only partially systematized and distributed over various Internet sites, portals and electronic archives.

To solve this problem, various Internet resources are developed. The Internet catalog LINGUIST List (http://linguistlist.org/), the most well-known of them, has been developed in order to provide a forum, where academic linguists could communicate and exchange information. It contains information about publications, persons, research organizations, grants, projects, foundations and conferences related to linguistic subjects.

Among other similar Internet resources, we should note Language Technology World (LT World) web portal (http://www.lt-world.org/) developed by the German Research Center for Artificial Intelligence (DFKI). The LT World data set contains information about technologies, resources, products and tools developed for natural language processing, and about persons, organizations and projects related to the Language Technology domain. Since this web portal is driven by ontology [1, 2], all data units mentioned above can be interconnected with each other by ontological relations. Unfortunately, this portal does not present information about Russian research projects.

As a Russian analog of LINGUIST List, we can consider a web portal “Linguistics in Russia: Resources for Researchers” (http://uisrussia.msu.ru/linguist/index.jsp), organized as a hierarchical catalog of references to resources related to computational, theoretical and applied linguistics, as well as a site RUSLING (http://rusling.narod.ru) including a catalog of references to vocabularies and text corpora for various languages and information about Russian linguists.

Internet resources describing particular branches of computational linguistics are also presented. As an example, we can cite a Russian site “Speech technologies” (http://speech-soft.ru/) which presents information about applied aspects of this branch (technologies, software tools, applied systems, systems project team, etc.), as well as the catalog of natural language generation (NLG) systems “John Bateman and Michael Zock’s List of NLG systems” (http://www.nlg-wiki.org/systems/) including information about practically all famous NLG systems (now the list contains description of 385 systems).

As a rule, almost all well-known linguistic Internet resources either are oriented to information support of linguistic communities and representation of common language information, or have narrow subject direction. And none of them is oriented to both integration of diverse resources on computational linguistics and support of content-based access to them for a wide circle of users.

To solve these problems, we have developed a specialized Internet portal – a knowledge portal on computational linguistics. As an information
resource, this portal provides representation of computational linguistics (CL) as a scientific discipline, data about researchers working in this direction, integration of resources on CL in a uniform information space, content-based access to systematized knowledge and data relating to CL, as well as information support of users and personification of user interface.

The users of this portal are researchers, lecturers and students involved in this branch of science and specialists who develop systems for natural language processing and speech analysis and synthesis.

2 Information model of knowledge portal on computational linguistics

In order to provide a user with the possibilities described above, the portal should not only have flexible means for presentation of heterogeneous information and content-based access to it, but also be easily adjustable to a new area of knowledge (branch of science) and provide operative management of its content.

To support a unified representation and storage of heterogeneous knowledge and information resources on CL and to provide a required functionality of knowledge portal, its informational model has been developed. This model joins the subject domain model with the problem domain model of the knowledge portal and describes the types of information presented in its content. Based on the information model, the portal internal data base is built and filling of the portal content, navigation through and search in the portal information space are organized.

The ontology is a core, the basic component of the information model of the portal. It not only describes the portal knowledge system, but defines the formal structure for representation of its content.

From the informal point of view, the portal ontology serves for representation of the concepts required for description of research activity and scientific knowledge in the whole and a specific branch of science “computational linguistics” in particular. Therefore, the portal ontology includes the universal ontologies of research activity and scientific knowledge [3], as well as the computational linguistics ontology.

The first two of the above mentioned ontologies are independent of the subject domain and can be used practically in any portal of scientific knowledge. Therefore these ontologies are selected as basic (see Fig.1). Let us consider them in detail.

The ontology of research activity is a top-level ontology and includes the basic concepts related to organization of research activity, such as Researcher, Organization, Event, Activity, and Publication. These concepts are used for describing participants of research activity, scientific events, research programs and projects, various types of publications and the materials represented in a printed or electronic format (e.g. monographs, articles, reports, proceedings of conferences, etc.). This ontology includes the concept Information resource that serves for describing resources presented in the Internet.

The ontology of scientific knowledge is in its essence a meta-ontology. It contains meta-concepts and relations which state the structures for describing the subject domain (branch of science) of the portal. In particular, this ontology includes meta-concepts, such as Subdivision of science, Subject of research, Object of research, Research method, and Scientific result. Using these meta-concepts, we can describe divisions and subdivisions that are significant for a given branch of science, determine classification of subjects, objects, methods of research, and describe results of research activity.

The concepts of the basic ontologies are interconnected with each other by associative relations (see Fig.1), a selection of which was performed taking into account not only preciseness and completeness of representation of the problem and subject domains of the portal, but also convenience of information search and navigation through its content.

The ontology of computational linguistics plays a role of the subject domain ontology of the portal. The 200 concepts of this ontology are realizations of meta-concepts of the ontology of scientific knowledge ordered in five “generic-specific” hierarchies (see Fig.1), each of which corresponds to one of the listed above meta-concepts. These hierarchies are Subjects of research hierarchy, Objects of research hierarchy, Subdivisions of science hierarchy, Research methods hierarchy, and Scientific results hierarchy. All hierarchies of computational linguistics are connected by means of associative relations. One part of these relations is inherited from the basic ontologies, the other part of them is specific relations of a given subject domain.

The subjects of research of CL are Processes and tasks related to functioning of units of language in the process of communication (Morphological analysis, Modeling of speech sound, etc.), and Applied processes and tasks which are of practical value and satisfy certain social needs. The Subjects of research hierarchy is connected with the Objects of research hierarchy by the associative relation “Aspect” and with Subdivisions of computational linguistics hierarchy by the relation “Subject for study”.
Fig. 1. The basic ontologies of the portal

*Nonverbal communications, Discourse, as an objective form of existence and application of a natural language, and Structural units of language corresponding to various levels of language, such as sentences, collocations, words, morphemes, sounds, etc., are considered as the basic objects of research in CL. The class of concepts *Discourse* is represented in the hierarchy by two subclasses Text and Speech. Structural units of language are grouped in compliance with the levels of language in classes, such as *Syntactic units, Lexical units, Morphological units, and Phonetic-phonological units*. To represent connections between entire discourse and their structural units, the relation “Inclusion” is used.*

The Research methods hierarchy serves for systematized description of the research tools that are applied in computational linguistics. Here the subclasses of concepts, such as *Methods of text processing, Methods of speech processing, Methods of analysis of text corpora, Methods of theoretical linguistics, Syntactic models, Logical models, Mathematical models and methods*, etc. are distinguished.

Subdivisions of the computational linguistics hierarchy is based on classification of the main theoretical directions of computational linguistics.

*Modeling of language and language activity, as well as Creation of applied systems, are considered as the main divisions of CL. The division *Modeling of language and language activity* includes subdivisions, such as *Automatic text processing (ATP), Text generation, Text recognition, Speech technologies (ST)*, etc. Other divisions contains subdivisions, such as *Creation of ATP applied system, Creation of ST applied system, Machine translation, Question-and-answer system*, etc. These general divisions are divided into more specific subdivisions. For example, *Machine translation includes Automatic and Automatized machine translation.*

The Scientific results hierarchy serves for structuring and description of results of research activity. It includes such classes as *Technologies and Software products, Applied systems and Linguistic resources*. The class *Linguistic resources* contains subclasses, such as *Dictionaries and Thesauri, Lexical Ontologies, Text and Speech Corpora*, as well as *Linguistic databases*.

Thus, introducing formal descriptions of concepts of a problem and subject domain in the form of concepts and relations between them, the portal ontology defines the structures for representation of real objects and connections between them.
According to this, the portal data themselves are represented as a set of interrelated heterogeneous information objects that in the aggregate constitute the portal content.

Formally, each information object (IO) corresponds to a certain class of ontology (is the instance of this class) and has a structure defined by this class. There can exist connections between information objects. Semantics of these connections is defined by relations between the corresponding classes of ontology.

3 Content of the knowledge portal

The portal content includes both a general knowledge represented by ontology and a specific knowledge about real objects and information resources systematized in compliance with the portal ontology.

As the portal is devoted to computational linguistics, its content presents, in the first place, knowledge about the main division of computational linguistics, its objects and subjects of research as well as models, methods and algorithms which are used in CL. Users of the portal can not only get the idea about computational linguistics as a scientific discipline, but also find information about research and production activity performed in this area. First of all, this is information about researchers, research teams and organizations and their activity.

Scientific and commercial projects, in the framework of which the great mass of linguistic knowledge and resources are produced, occupy a special place in the activity of research organizations and researchers. Results of this activity are presented in publications, i.e. monographs, articles, reports, proceedings of conferences and other text resources, access to which is provided by the portal.

Besides, the portal provides access to information resources presenting the direct results of activity of organizations and individual researchers obtained within scientific and commercial projects. Such results are the technologies, software products, applied systems and traditional linguistic resources, such as vocabularies, text and speech corpora, and linguistic databases.

To make access to resources describing these results more efficient, the portal content presents information about various aspects of their development (acquisition), such as organizations, researchers and projects concerned with their obtaining, as well as information about such substantial properties of the results of activity as reference (or relation) to the division of science, research method, object or subject of research. This information links these resources to the rest of knowledge and data presented in the portal content.

It allows a user to select groups of resources produced, for example, during a certain research activity (e.g. grant, project) or using a certain class of research methods.

Description of Internet resources systematized in compliance with the portal ontology is an important component of the portal content. Such resources are sites of organizations, conferences, projects, portals and catalogs related to computational linguistics, as well as personal Internet pages of researchers.

According to the definition of the basic ontologies, each Internet resource corresponds to the concept of Information resource. The set of attributes and relations of Information resource is based on Dublin Core standard (http://dublincore.org/documents/usageguide/) and includes the following units: “Title of the resource”, “Address in the Internet” (URL), “Subject of the resource”, “Resource type”, “Language”, etc. The description of the resource includes an instance of this concept (an information object) and a set of instances of relations that allow one to link a given resource with researchers, organizations, events, projects, results of researches, etc.

4 Adjustment of the knowledge portal and management of its content

Adjustment of the portal to a certain subject domain (area of knowledge) and user preferences and management of its content are performed with the help of specialized editors (ontology, thesaurus and data editors) implemented as web applications and accessible to authorized users through Internet (see Fig. 2).

Adjustment of the portal to a given subject domain is realized with the help of the ontology editor that allows one to create, modify and delete any elements of the portal ontology (classes, relations, domains, constraints) as well as to define and modify the concept hierarchies.

The ontology editor was designed in such a way that it is easy to understand and use for experts who are not experienced in programming and computer science. In particular, to meet these requirements, we refused to use such popular means of ontology building as the Protégé editor (http://protege.stanford.edu/).

To make representation of information more convenient for a user, the facility for adjustment of knowledge and data visualization is provided by the ontology editor. This facility allows one to define for each class of the portal ontology the template of visualization of information about its objects and template of visualization of semantic (meaningful) references to its objects.
In particular, the template of visualization of semantic reference to the class objects can include both attributes of its class and attributes of classes that are linked with it by relations and attributes of these relations. The values of attributes included in this template are used for building the text presentation of a hyperlink to a class object.

To adjust the portal to various languages, the multilingual thesaurus [4] is included in the portal. The thesaurus is built as a linguistic complement of ontology and contains the terms of the problem and subject domains of the portal, i.e. words and phrases in several natural languages by means of which the concepts and relations of ontology are presented in texts and user queries. Thereby, as opposed to ontology where concepts are connected by relations typical for a given area of knowledge, in the thesaurus terms are connected by traditional linguistic relations, such as synonymy, equivalence, etc.

Relations defined between the multilingual thesaurus terms and the ontology concepts provide visualization of the portal ontology and the information presented in the portal content in different languages (at user option) and support navigation through its content and forming a retrieval query in a natural language suitable for a user.

Management of the portal content is performed with the help of the data editor that is ontology-driven. It allows one to create, modify and delete information objects and relations between them. Forms for input of such data are automatically generated on the basis of classes and relations of the portal ontology.

To make automatic extension of the portal content with information about Internet resources relevant to the portal area of knowledge, a collector of ontology information about Internet resources has been developed. It performs search, collection, analysis, evaluation of relevancy of Internet resources, as well as its automatic indexing and classification.

The indexing and classification of resources are performed using ontology and thesaurus. As a result, the semantic index (semantic annotation) is built for each resource, and the subdivision of CL to which it refers is defined. The semantic index of a resource contains a set of objects and relations that present its content in terms of the portal ontology. For each resource an information object – an instance of the class Information resource – is built and linked with all objects included in its semantic index. All these objects and their relations are inserted in the portal content and become accessible for the search and navigation engine.
5 Providing a content-based access to the portal content

A content-based access to the systematized knowledge and information resources on CL presented in the portal are realized by means of both the navigation through a tree of the ontology concepts and the portal content (along with associative relations) and the advanced facilities for semantic search (using concepts and relations the portal ontology).

5.1 Navigation through the portal content

As said above, for a user of the portal the data are presented as a set of interrelated information objects. During navigation through the portal content, there is a possibility of choice of information objects of a required concept (class), browsing and filtration of the list of such objects, detailed browsing of each IO and transition along the ontological links (relations) from this IO to other information objects, as well as browsing of Internet resources the references to which are contained in this IO.

Navigation through the portal content starts with the choice of a certain class in the tree of ontology concepts and the portal content (along with associative relations) and the advanced facilities for semantic search (using concepts and relations the portal ontology).

![Fig.3. Presentation of an information object and its connections](image)

Fig.3. Presentation of an information object and its connections
the user can proceed to their (objects) detailed description.

Further navigation through the portal content is a process of transition from one information object to others along the ontological links (instances of associative relations) defined between them.

For example, when we look through information on a concrete project (e.g. AGILE), we can see both values of its attributes and its connections with other objects (see Fig.3). Using these connections as elements of navigation, we can proceed to browsing of detailed information about the project using both direct relations (about the object of research, scientific results and methods which are used in the project, results obtained in the project) and inverse relations (about participants of the project (persons and organization), publications and Internet resources describing this project).

When we move along a certain link of any information object, we can get a rather big list of objects, for example, a list of publications of a well-known scientist. To solve this problem, the procedure of lists filtration has been implemented.

The filter is a set of conditions that define admissible values of attributes of information objects and requirements to existence of connections with other IOs. This method allows us to filter a set of publications by date (condition on an attribute), by scientific result described in this publication or by object of research (conditions on an object connected with a given object).

5.2 Information search in terms of the subject domain of the portal

When a user searches for certain information, he can formulate his query in terms of the portal subject domain. For that he must select a class of the sought object and define constraints on the values of attributes of the sought object and on its relations with the other objects. A user can also define constraints on the values of attributes of objects connected with the sought object by associative relations.

For example, the query “Find the research methods used for processing of business letters in Russian for the period from 1998 to 2005” will be presented as follows:

Class “Research method”:
Relation “Applied to object”:
Class “Object of research”
Attribute “Name of object” = “Business letter”
Attribute “Language” = “Russian”
Relation “Uses research method”:
Class “Project”

The retrieval queries are formed by means of a special graphic interface driven by the portal ontology. When a user selects a class of the sought objects, a retrieval form is generated automatically, where the user can define constraints on the values of attributes of the sought object, as well as on the values of attributes of objects connected with the sought one.

6 Conclusion

The paper describes a knowledge Internet portal that presents the systematized knowledge about the main divisions of computational linguistics, its subjects and objects of research, as well as models, methods and algorithms which are used in CL. It contains data about researchers, research teams and organizations involved in research on CL, and about projects performed in this area.

Users of the portal also have a content-based access to information resources presenting the actual applied systems, technologies and software products, intended for natural language processing, as well as linguistic resources and databases.

Because systematization and structuring of knowledge and information resources on CL are performed on the basis of ontology, access to them is realized by means of navigation through the tree of the ontology concepts and the portal content (along with relations presented in the portal ontology), as well as using the facilities of semantic search.

To create the knowledge portal on computational linguistics, we have used the tools, methodology and technology suggested in [3, 5].

Now the knowledge portal is accessible at http://uniserv.iis.nsk.su/cl. At the moment, a representative core of the ontology on computational linguistics including about 200 basic concepts is developed. The portal content includes more than 600 Internet resources, about 2000 information objects linked by more than 4000 relations.

Our immediate goals are improvement and elaboration of the ontology on computational linguistics and the multilingual thesaurus linked with it, as well as extension of the portal content with new linguistic resources.

Besides, we plan to include the facilities of graphical visualization in the portal. It will allow us to present in the form of a graph not only the hierarchy of the ontology concepts, but also the portal content.
7 Acknowledgments
The authors are grateful to the Russian Foundation for Basic Research (grant № 09-07-00400) for financial support of this work.

References: