DEVELOPMENT OF THE ALGORITHMIC SOFTWARE AND INFORMATION SUPPLIES FOR THE LINGUISTIC ONTOLOGY BASED ON STRUCTURED ELECTRONIC ENCYCLOPEDIC RESOURCE FORMATION METHOD

Anton Mykhailiuk¹, Vasyl Teslyuk²
¹Borys Grichenko Kyiv University, Kyiv, Ukraine; ²Lviv Polytechnic National University, CAD Department, Lviv, Ukraine

Abstract. Intellectualization of the information retrieval process requires appliance of the specialized linguistic resources: The domain specific linguistic ontology might be one of such resources. This article presents software organization method for automated formation of the ontological knowledge base by converting structured encyclopedic resource to the appropriate ontology objects. The concepts ontology database creation procedures, concepts hierarchies and associative relations networks are presented and also studies of the qualitative and quantitative composition of the current experimental ontology based on Ukrainian Wikipedia segment are performed.

Keywords: linguistic ontology, semantic relations, structured encyclopaedia

1. The linguistic ontologies features analysis

Based on the references analysis, the following important features for structurally-logical characteristics of similar ontologies can be highlighted:

- different concepts can have the same linguistic representation, in this case, it is talked about disambiguas terms and semantic differentiation of such concepts is possible by using interpretations;
- all synsets are combined into a single hierarchical terms taxonomy from abstract concepts to concrete ones;
- ontology synsets are linked together by semantic relations, the most common relations are association and hyponym-hyperonym [2].

In addition, the ontologies use in automatic mode for synthetic languages, which include Ukrainian language, requires normal word forms for ontology concepts.

An ontology formation by software means involves separate text data analysis, concepts separation and identification of links among them in accordance with the specified semantic relations. Unfortunately, this procedure requires too complicated semantic analysis apparatus, which, however, does not guarantee the ontology quality due to natural texts interpretation ambiguity. The solution to this problem lies in bringing the collection documents of encyclopedic (vocabulary) type, which have a clear structure and are formed by experts. The free electronic encyclopedia Wikipedia [14] is an example of such structured collection. Its articles have concepts descriptions, moreover the narrative part has quite clear structure and, in addition, contains links to other articles. This enables automation of such documents processing and an ontology development.

Therefore, the purpose of the presented study is to develop ways of software organization, which would enable automatic or semi-automatic mode for creation a linguistic ontology knowledge base on bases of the certain language segment Wikipedia articles (e.g., Ukrainian segment) for application in retrieval process, in particular, in quasiharmonic search procedures [9].

2. Development of the input data structure

The ontology development software has to use the structural organization of Wikipedia materials features. The articles are formed with the use of a special format MediaWiki [5] (fig. 1). This format allows unique description of the structural elements (headers, sections, links, meta elements, etc.) and their efficient processing in automatic mode. The entire archive collection for a particular language Wikipedia articles segment is stored in the
form of an XML document [3, 4, 5, 10, 11, 13] and is available for downloading. The XML markup allows selecting necessary components for further processing. Fig. 1 shows an example of a node that describes one of the Ukrainian segment articles. This fragment illustrates the XML node structure for the article describing “Tuple” concept. The node contains not only the article source code, but also some additional meta-information. <title> (article title), and <text> (text of the article) fields are of the greatest interest for the content analysis and the problem solving.

![Image](71x489 to 249x680)

**Fig. 1. The Wikipedia article XML document fragment**

Table 1. The types of Wikipedia articles

<table>
<thead>
<tr>
<th>Articles type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory articles</td>
<td>Articles describing a particular concept, event or phenomenon. Accordingly to the title is the main source of the encyclopedia information content and serve as the main source for the ontologies.</td>
</tr>
<tr>
<td>Articles describing the multi-valued concept</td>
<td>This kind of articles is designed to store a list of all currently available in the corresponding Wikipedia segment means of some term. The following articles contain links to the appropriate explanatory article, if it is available, and a short, unique description for each semantic meanings of the term. The main indicator of such articles is the proprietary tags: [disambig] at the beginning of the text.</td>
</tr>
<tr>
<td>Categories articles</td>
<td>Article describing the concept-category from the general Wikipedia categories hierarchy. In the explanatory articles, one or more of such concepts can also be specified as a parent category.</td>
</tr>
<tr>
<td>Articles describing the files</td>
<td>File articles describe the Wikipedia file objects (e.g., images) and contain information specific for the respective objects (references, size, type, etc.)</td>
</tr>
<tr>
<td>Incomplete articles</td>
<td>Explanatory articles often link to articles, which due to some reasons, are not yet written. They do not have the narrative part, but have the title.</td>
</tr>
<tr>
<td>Service articles</td>
<td>Such articles do not have direct informational value for the user, and are used in the Wikipedia development process. In particular they might be article templates, helps, comments, discussions, etc.</td>
</tr>
</tbody>
</table>

The entire set of Wikipedia articles by various criteria can be attributed to several groups which are depicted in the table. 1. Analysis of the Wikipedia structure allows concluding that Wikipedia articles along with reciprocal internal links create a prototype of ontological knowledge bases [14]. The main role in this case is played by the explanatory articles and categories article, the multivalued concepts articles group performs a supporting role in development of the explanatory articles for each of the meaning. In addition, the incomplete articles can be used to identify relations among other articles if those contain the according link.

3. Development of the output data information model

The detected Wikipedia structure features allow examining the software for linguistic ontology creation as means of Wikipedia elements conversion into the appropriate ontologies objects. Since the last will be actively used in retrieval procedure, their storage with an effective and prompt access to them can be resolved by applying the databases system. According to the logical organization of the ontology, fig. 2 shows the designed structural scheme of a relational database that contains the ontologies main components and relations among them. The result of the ontology information content formation would be filled appropriate values of the database tables. Lets consider in details all the relational links in the database. Synset is a central relational relation which reflects the ontology synset. It contains the following attributes: a unique identifier (id), a symbolic representation of a sinset in Ukrainian (ua), Russian (ru) and English (en), a unique semantic interpretation of a synset (descr). The ru and en fields are introduced to establish a potential ontologies connection with the similar ontologies, created on the basis of Russian and English Wikipedia segments. Depending on the requirements, a set of languages can be enhanced by adding the appropriate fields in the synset relational relation or completely removed from additional language attributes, leaving only the main language name synset.

![Image](71x489 to 249x680)

**Fig. 2. The ontologies database structure**

Vocab is an auxiliary relational link designed to store all synonymic synset inputs. It consists of a unique identifier (id), a term, which is a symbolic representation of synonymic input (term), an identifier of the parent synset (synset_id) and a field that contains number of words present in term (parts). Norm is an auxiliary relational link for storing word normal forms that make up the term symbolic representation from Vocab. It consists of a unique identifier (id), a character based representation of a word from a term normal form (word) and the identifier of the parent term from Vocab (vocab_id). Relation is the keyword relational link that reflects existing relations among the ontology concepts. The synset_id1 and synset_id2 fields contain synsets identifiers, between which there is a semantic relations. The weight filed characterizes this relationship by a certain weight value, the higher the value, the stronger is the relations between synsets. Finally, the type field stores the semantic relation type, thus, in the quasiharmonic search context there are two types of relations — association and hyponym-hyperonym.

Such source data structural organization allows forming a hierarchically structured ontology knowledge base for specific domain field, which will describe the concepts (the concept) of outer world and a variety of relations linking these concepts. This knowledge base can be used as a semantic core for quasiharmonic search process. The quasiharmonic approach will allow including semantics in nonsemantical search engines through the use of semantic coherence notion that is embedded in the ontological knowledge base, during the search query formation. To modify the search query within the quasiharmonic search it is proposed to use the knowledge base for domain field via three directions. The first direction - fixing terms from the search query using certain ontological knowledge base notions: the terms are allocated with the use the normal words forms (norm) table, and with the help of vocab table they are bind with a certain synset in case of one valued relation, or the of homonymy resolving procedure for a given term, in the case of its ambiguity is initiated. Fig. 3 shows the database query diagram, which allows determining the corresponding synsets for the normalized words from a user search query (e.g. word1, word2, etc.).
Fig. 3. The database query scheme for searching synsets, which match the words from the search query.

Fig. 4. The database query scheme for subsidiary and parental elements search.

Fig. 5. The database query scheme for searching the ontology semantic cut.
In this way the implicit search query modification is done, as it goes from the words set plane into the concepts plane.

The second way is request modification by refinement or enlargement of the query. It is done due to synonymous concepts groups usage, as each concept from the synsets table (synset) can have multiple synonymous representations in the form of terms from a terms table (vocab). Modification may also be done due to concepts relations through common parent elements in the ontology hierarchical structure, and also due to parental and subsidiary concepts. Fig. 4 shows the database query diagram for the according subsidiary and parental synsets search for the selected synset (<synset_id>).

It is possible to move focus from the general search results to more vague and vice versa by navigating the ontology vertical ties from the relation table, and inclusion into the query concepts logically associated with the the initial query version terms, but more or less specific from them depending on the user's information interest.

Finally, the third direction is the request modification by the search direction management through the semantically linked environment of the current query concepts set. The mechanism for automatic search and visualization of the domain specific ontology sections, that are mostly gravitated to the query concepts group, enables nonlinear (i.e. by hidden associative relations) modifying the request for clarification or change of the search vector. For this aim in the domain field ontology, with the use of the relation semantic table, subnets of the specific radius are determined, each of them has the according query notion in its center. Fig. 5 shows the database query diagram for searching the first (round1) and the second (round2) synsets rounds from some syns subnet <synset_id>. The zero round (round0) is entered by the synset and the temporary table round0 round1 is intended to exclude synsets, which already are included in the previous rounds from the second round (round2).

The intersection of these subnets, taking into account the weight (weight) generates the ontology cut by request concepts and can point out the ontological s areas that are semantically most gravitated towards the user's search interests

4. The algorithmic organization of the ontology creation

The ontology formation process can be divided into several stages, laid into the basis of the corresponding algorithm:

- preparatory phase;
- stage of database synsets creation;
- stage of hierarchical relations construction;
- stage of associative relations construction;
- correction phase.

In accordance to the presented general algorithm for ontology creation the process begins with the XML data parsing and the corresponding synset base formation and completes with implementation of the corresponding correction of the formed ontology database in the automatic or manual mode if one is needed.

The developed algorithm has sufficient flexibility of the ontology formation process, in the same providing simplicity and reliability of its programming implementation and possibility of further functionality enlargement by introducing additional methods.

The presented in the chart (Fig. 6) class mediawiki_parser provides all the above steps for the ontology creation. Since the source data of the Wikipedia articles are available in XML format, the initial phase with the main objective of XML-data parsing and their preparation for further processing is required. According to the data segment presented above, it is essential to navigate to each subsequent article (page node) in the ontologies creation process. The title (title node) and the text of the article (text node) from a page node subtree are of further interest for further ontology creation.

---

The presented general algorithm for ontology creation algorithm is intended to exclude synsets, of which only are reselected into the temporary table round0 round1, to whose article it is refereed;

The intersection of these subnets, taking into account the weight (weight) generates the ontology cut by request concepts and can point out the ontological s areas that are semantically most gravitated towards the user's search interests

The algorithmic organization of the ontology creation

The ontology formation process can be divided into several stages, laid into the basis of the corresponding algorithm:

- preparatory phase;
- stage of database synsets creation;
- stage of hierarchical relations construction;
- stage of associative relations construction;
- correction phase.

In accordance to the presented general algorithm for ontology creation the process begins with the XML data parsing and the corresponding synset base formation and completes with implementation of the corresponding correction of the formed ontology database in the automatic or manual mode if one is needed.

The developed algorithm has sufficient flexibility of the ontology formation process, in the same time providing simplicity and reliability of its programming implementation and possibility of further functionality enlargement by introducing additional methods.

The presented in the chart (Fig. 6) class mediawiki_parser provides all the above steps for the ontology creation. Since the source data of the Wikipedia articles are available in XML format, the initial phase with the main objective of XML-data parsing and their preparation for further processing is required. According to the data segment presented above, it is essential to navigate to each subsequent article (page node) in the ontologies creation process. The title (title node) and the text of the article (text node) from a page node subtree are of further interest for further ontology creation.

---

The instance of mediawiki_page class for work with the Wikipedia page in the MediaWiki format (fig. 6) is created on the basis of the received data. It allows manipulating the article data, doing various checks, selecting logical sections, etc.
method, which, in its turn, automatically triggers the _add_norm method. Fig. 8 presents the pseudocode, which abstractly shows the sequence of actions at this stage (_fill_synset method of the mediawiki_parser class).

```java
/* Create XMLReader */
xml = new XMLReader(config_file);
/* Form synsets database */
/* Read the next page node, while it is possible */
while (xml_page = xml.next_page()) do
  /* Create article object */
  page = new mediawiki_page(xml_page->get_title(), xml_page->get_text());
  /* Launch synsets filling method based on the current article */
  this->fill_synset(page);
end while;
/* Return the XMLReader indicator at the beginning of the file */
xml->reset();
/* Form ontology relations */
/* Read the next page node, while it is possible */
while (xml_page = xml.next_page()) do
  /* Create article object */
  page = new mediawiki_page(xml_page->get_title(), xml_page->get_text());
  /* Launch hierarchical relations filling method */
  this->fill_hierarchy(page);
  /* Start the filling method of associative connections */
  this->fill_assoc(page);
end while;
```

Fig. 7. Pseudocode for the ontology database filling

The ontology hierarchical relations creation stage, as it was noted earlier, is based on the use of Wikipedia articles hierarchies that describe the category. At the end of each explanatory article or one describing the category the list of parental categories to which this article belongs is indicated. It is easy to notice that within the Wikipedia one concept may refer to several categories at the same time.

```java
/* If article does not describe synset */
if page->is_not_synset() then
  return; /* stop work */
/* If article describes reference to another one */
else if page->is_redirect() then
  /* Add new synset with symbolic presentation of the concept from the article */
  /* to which it is redirected, the current concept is given as a synonym */
  this->add_synset(page->get_redirect_synset(), page->get_synset());
/* If article describes multi-valued concept */
else if page->is_disambigous() then
  /* Get list of all synsets and their explanations */
  list_synset_description = page->get_disambigoussynsets();
  /* Create new synset for each entry from the list */
  for each record in list_synset_description do
    /* Add new synset (synset) with the definition (description) */
    /* the current article title gives as a synonym */
    this->add_synset(synset, description, page->get_synset());
  end for;
/* If article describes concept or category */
else
  /* Add new synset */
  this->add_synset(page->get_long_link(), page->get_long_link(en), page->get_description());
end if;
```

Fig. 8. Pseudocode for the ontology/synsets base filling

The hierarchical structure formed in this way will differ from the classic terms hierarchy, where each term has only one more general concept (in other words, only one parent). However, this limitation does not correspond to the real relations among concepts, since for most of them it is impossible to uniquely detect only one parent. Therefore, the direct conversion of all Wikipedia categorical links into the hierarchical structure of linguistic ontology synsets is regarded as the most optimal, provided that during the design of toolkits, which will use direct hyperonymy relations, the above feature of such relations formation will be taken into account. According to this pseudocode the _fill_hierarchy method will have the structure shown in fig. 9.

```java
/* If article does not describe concept or category */
if page->is_not_synset() OR page->is_redirect() OR page->is_disambigous() then
  return; /* stop work */
/* If article describes concept or category */
else
  /* For each article category */
  for each category in page->get_categories() do
    /* Save the hierarchy relation with relation weight equal to w */
    database->relation->insert(
      synset_id = category,
      synset_id = page->get_synset(),
      type = H,
      weight = w);
  end for;
end if;
```

Fig. 9. Pseudocode for ontology hierarchical relations formation

The associative relations formation stage fundamentally differs from the hierarchical relations formation stage only by the fact that here for relations identification the links to other Wikipedia articles (and, consequently, other concepts in ontologies) in the current article text are used. Moreover, since the link can be repeated, and there also might be reverse links (when the article to which one was referred contains a backlink), this allows increasing this relation weight by some delta value, thereby reflecting the semantic relation strengthening. The higher number of such links, the stronger the relationship among articles, respectively, the greater must be the weight ratio. In order to distribute the relations weight in the ontology uniformly enough, delta should be chosen relatively small in accordance with the initial weight value w0. The associative relations are filled in the _ill_assoc method, its pseudocode can be seen in fig. 10.

Finally, the last step is an optional correction phase of already created in the previous stages ontological base. It can be done automatically (for example, removing possible service concepts and their relations), and by manual editing of the ontological base by ontology editors (for example, removal of irrelevant or erroneous relations etc.). This correction may be helpful due to partial transparency of the automatic ontology formation and the availability of proprietary information in the primary source, which can be displayed in a specific encyclopedia articles (e.g., category "article starting from the letter "A", etc.), which should not be included in the ontology.

5. Results and conclusions

For creation and analysis of an experimental linguistic ontology the research prototype of the Ukrainian Wikipedia segment into the appropriate Ukrainian ontological knowledge base conversion software has been developed. The results of the software module for linguistic ontology based on the Ukrainian Wikipedia segment articles detected during the experiment are included in the table 2. The observed quantitative characteristics allow arguing that this ontology covers quite wide area of human knowledge and can be used in information systems. It should also be mentioned that the developed ontology is only the initial phase, the basis for the further work for experts and, at the same time, scientific-information resource for research and experiments in the search technology field. In further work with the developed ontology, it is possible to provide new concepts and links or remove the false ones.
The qualitative analysis of the developed ontologies is, in general, actually impossible, taking into account very large number of concepts and relations among them. This process requires long-term practical use of the ontology and considerable human and time resources. Improvement of the ontological knowledge base quality must be performed in parallel with its usage in the development of information retrieval, content analysis procedures, etc. However, the estimation can be done by means of selective analysis. For this aim several small scatter subnets (such as shown in fig. 11) from different parts of the ontology are build and assess the major links among synsets.

Analysis of these fragments allows concluding that concepts and relations among them, with some minor exceptions, mostly correspond to the conceptual structure of the subject areas. Thus, qualitative composition of the ontology and adequacy of the relations can be considered quite acceptable for use in information retrieval field.

Thus, the presented ontology formation procedures enable relatively easy automatic development of quite large linguistic ontology, which would meet the basic requirements to such resource.

In addition, since Wikipedia as the primary source of information for ontology development is multilingual, in perspective it is possible to create cross-language ontologies, which would greatly expand the scope of their application.

**Table 2. Quantitative characteristics of the ontology**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Quantitative value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synsets number</td>
<td>243948</td>
</tr>
<tr>
<td>Average number of words in synset</td>
<td>0.15</td>
</tr>
<tr>
<td>Number of hierarchical relations</td>
<td>127366</td>
</tr>
<tr>
<td>Number of associative relations</td>
<td>2658584</td>
</tr>
</tbody>
</table>

**Literature**