An Approach to the Selection of DSL Based on Corpus of Domain-Specific Documents

E. Elokhov, E. Uzunova, M. Valeev, A. Yugov, V. Lanin
Department of Business Informatics
National Research University Higher School of Economics
Perm, Russian Federation
eugene.yelokhov@gmail.com, palgonuri@gmail.com, mt.vallev.1992@gmail.com, yugovas@live.ru, lanin@perm.ru

Abstract. **Today many problems that are dedicated to a particular problem domain can be solved using DSL. Thus to use DSL it must be created or it can be selected from existing ones. Creating a completely new DSL in most cases requires high financial and time costs. Selecting an appropriate existing DSL is an intensive task because such actions like walking through every DSL and deciding if current DSL can handle the problem are done manually. This problem appears because there are no DSL repository and no tools for matching suitable DSL with specific task. This paper observes an approach for implementing an automated detection of requirements for DSL (ontology-based structure) and automated DSL matching for specific task.**

Keywords: ontologies, conceptual search, domain-specific language, semantic similarity of words

I. INTRODUCTION

Nowadays metamodeling and DSL-based technologies (DSL – Domain Specific Language) [16] are widely used in information system developing. DSL is created for solving some specific problem. Almost every arising problem is similar to the one that was solved before. In this case it means that a suitable DSL was already implemented or an implemented DSL does not fully meet the requirements. Therefore, you can either find a ready-to-use DSL or complete and configure a DSL implemented earlier. This requires less costs rather than developing a completely new DSL.

So, there are two steps to select one of already existing DSL:

1. Determine the requirements for DSL.
2. Find out how closely each of DSL meets this requirements.

Requirements are determined by analyzing domain-specific documents or problem statement. Then a requirements ontology based on that analysis is generated.

To match a concrete DSL with generated ontology some matching metrics and DSL description formats must be defined. In this work the MetaLanguage system [1] allowing DSL creation will be used. The use of MetaLanguage system is justified by its noticeable features:

1) *the ability to work with most common DSL notations;*
2) *DSL convertation from one notation to another;*
3) *exporting dsls to external systems.*

In summary, the input data will be:

- corpus of domain-specific documents;
- set of DSL descriptions.

The target output is a list (ordered by correspondence to the generated ontology) of appropriate DSLs that can handle the problem.

This paper shows generating process of requirements ontology based on domain-specific documents and how a particular DSL meets given requirements.

II. RELATED WORKS

Nowadays there are some information systems that let you create text-based ontology models of documents or let you define correspondence of ontology models thereby transform one model onto another one. We found two web-resources that let you create ontologies: OwlExporter and OntoGrid.

The core idea of OwlExporter is to take the annotations generated by an NLP pipeline and provide for a simple means of establishing a mapping between NLP (Natural Language Processing) and domain annotations on one hand and the concepts and relations of an existing NLP and domain-specific ontology on the other hand. The former can then be automatically exported to the ontology in form of individuals and the latter as data type or object properties [7].

The resulting, populated ontology can then be used within any ontology-enabled tool for further querying, reasoning, visualization, or other processing.

OntoGrid is an instrumental system for automation of creating domain ontology using Grid-technologies and text analysis in natural language [12].

This paper is supported by Russian Foundation for Basic Research
(Grant 12-07-00763)

This system has bilingual linguistic processor for retrieving data from text in natural language. Worth D. derivational dictionary is used as a base for morphological analysis [4]. It contains more than 3.2 million word forms. The index-linking process consists of 200 rules. “Key dictionary” is determined by words allocation analysis in text. The developers came up with new approach of revealing super phrase unities that consist of specific lexical units. The building of semantic net is carried out this way: the text is analyzed using text analysis system, semantic Q-nets are used as formal description of text meaning [18]. The linguistic knowledge base of text analysis system is set of simple and complex word-groups of the domain. This base can be divided into simple-relation-realization base and critical-fragment-set, that let you determine which ontology elements are considered in this text. The next step is to create and develop the ontology in the context of GRID-net. A well-known OWL-standard is used to draw the ontology structure.

Also three information systems were found that fulfill a function of transformation [10].

ATLAS Transformation Language is a part of the architecture of managing ATLAS model [6]. ATL is the language that let you describe initial model transformation into destination model.

GReAT (Graph Rewriting And Transformation) is the language of model transformation description, which is based on triple graph transformation method [4]. This transformation represents the set of graph sorted re-record rules that are applied to the initial model and as a result create the destination model.

VIATRA is pattern-based transformation language for graph models managing which combines two methods: mathematic formal description (based on graph transformation rules for model description) and abstract finite state automaton (for control flow description) [5].

The program resources described before are key functions that determine an appropriate DSL matching. Unfortunately, a software system, which implements all this functions, was not found.

In addition, the idea used in applications intended to transform the ontology can be implemented to determine the measure of DSL correspondence to ontology requirements.

III. Approach description

The suggested approach of the DSL selection process consists of six stages that can be described as a series of sequential operations which should be implemented (fig. 1).

Firstly, a corpus of documents is processed. As a result, the key words (concepts related to specific domain) are retrieved. Secondly, when re-viewing the document, the relations between concepts are built. These concepts and relations form a semantic network. The next step is to eliminate synonymy (to merge nodes containing synonymic concepts). In order to achieve this, a linguistic ontology is used. After that, it is necessary to transform “contracted” semantic network into ontology model, using the graph coarsening algorithm with implementing linguistic ontologies. The next step is to qualify the ontology model by a specialist. This step includes concepts editing and relations marking semantically.

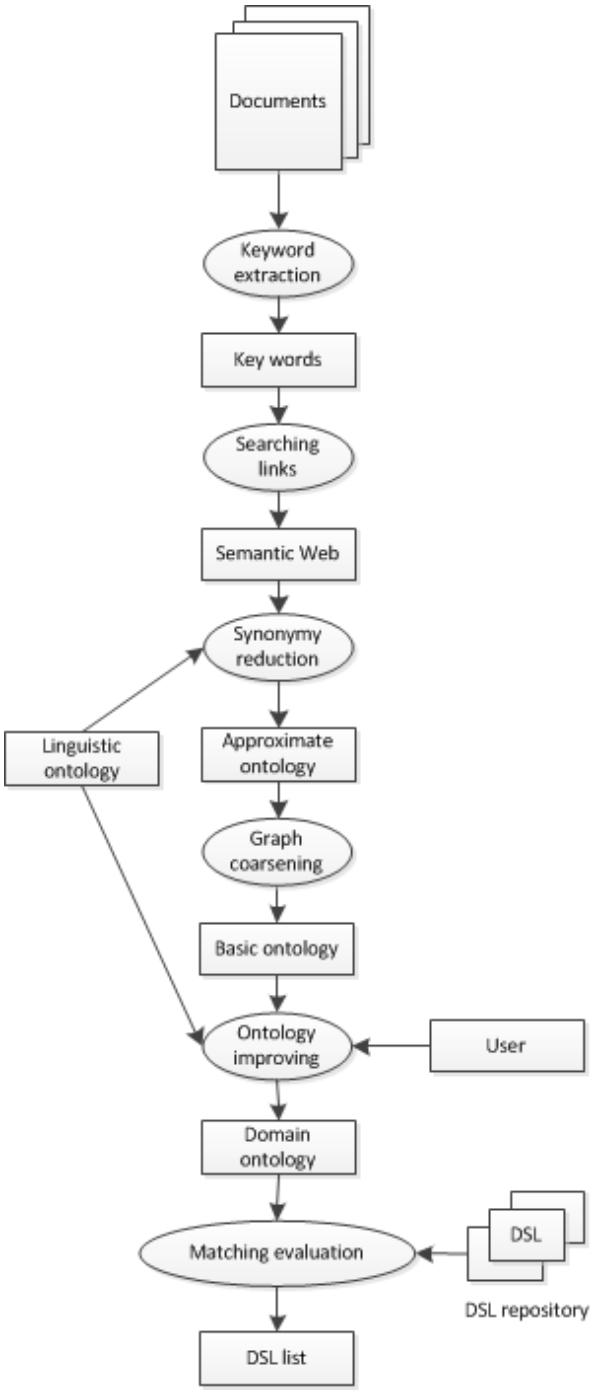

Figure 1. DSL selection process stages

When the ontology is complete, i.e. it meets user requirements, DSLs are taken from the repository, and the measures of DSLs correspondence to ontology requirements are calculated.

A. *Keyword extraction*

Using ontology is one of the most widespread ways to structure information on domain [11]. The formal ontology description is $O = <X, R, F>$, where

- X – a finite set of domain terms,
- R – a finite set of relations between the terms,
- F – a finite set of interpretation functions.

Within the context of this paper, let us take a look at defining the set of terms and the set of relations.

Consider that basic terms in document are its key words-nouns. Researches related to finding key words in documents are based on frequency laws discovered by linguist and philosopher George Kingsley Zipf. The first law says that multiplication of word detection possibility and frequency rank is constant. The second law says that frequency and number of words with this frequency also have a relation.

Currently, for searching key words the pure Zipf's laws (TF-IDF) and also LSI (latent semantic indexing) algorithms are used. This research observes Zipf's laws, which are easily implemented, and a linguistic processing will be provided by program resources of Aot.ru.

As an example some university exam taking process is described. Consider that frequency analysis retrieved following keywords (fig. 2).

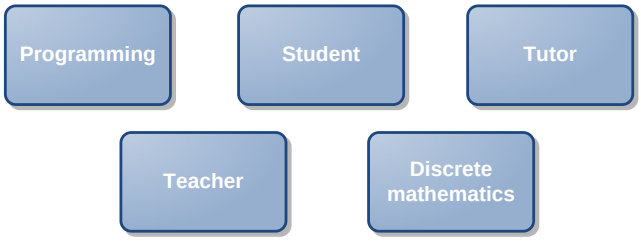

Figure 2. Exam taking keywords

B. *Searching relations*

As a result of frequency-response analysis we have a set of unlinked nodes (fig. 1). Now we have to define a set of relations, in other words to make disconnected graph a semantic net.

Semantic graph is weighted; its nodes are the terms of analyzed documents. The existence of edge between two nodes means that two terms are related semantically; weight of the edge is measure of semantic similarity [17].

Similarity measurement of ontology concepts can be calculated as follows:

1. Jaccard similarity coefficient [8]:

$$K_J = \frac{c}{a+b-c}$$

It's a statistic used for comparing the similarity and diversity of sample sets, where a – frequency of occurrence of first term, b – frequency of occurrence of second term, c – frequency of occurrence of joint terms.

2. Mutual information [2]:

$$MI = \sum_{u=\{0,1\}}\sum_{v=\{0,1\}} P(u,v)\log_2\frac{P(u,v)}{P(u)P(v)} \approx \sum_{u=\{0,1\}}\sum_{v=\{0,1\}} \frac{(u,v)}{N}\log_2\frac{(u,v)}{(u)(v)}N$$

where u, v – terms retrieved from the document; (u) – frequency of occurrence of u, (v) – frequency of occurrence of v, (u,v) – frequency of occurrence of joint u and v.

Point mutual information may be calculated as [2]:

$$PMI(u,v) = p\left(\frac{(u,v)}{p(u)\,p(v)}\right).$$

After calculating measurements of ontology concepts they must be averaged [15]. Based on average measurement, keywords become connected. As a result the semantic net (fig. 3) is created.

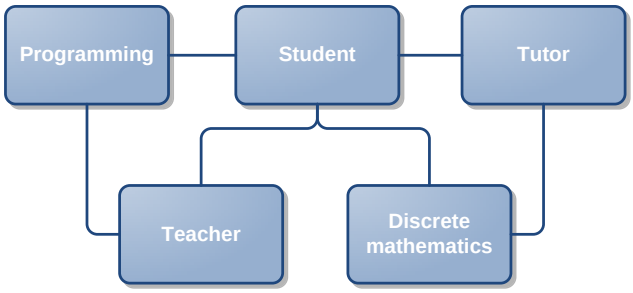

Figure 3. Exam taking semantic network

C. *Synonymy reduction*

Each concept is searched in linguistic ontology and those marked as synonyms are being contracted to a single node.

We are going to use *WordNet*, the semantic net, which was created at the Cognitive Science Laboratory of Princeton University. Its dictionary consists of four nets: nouns, verbs, adjectives and adverbs because they follow different grammatical rules. The basic dictionary unit is *synset*, combining words with similar meaning. It is also the node of the net. Synsets may have a few semantic relations like: *hypernym* (breakfast → eating), *hyponym* (eating → dinner), *has-member* (faculty → professor), *member-of* (pilot → crew team), *meronym* (table → foot), *antonym* (leader → follower). Different algorithms are widely used, for instance, the ones that take into account the distance between conceptual categories of words and hierarchical structure of *WordNet* ontology.

Linguistic ontology showed that example's *tutor* and *teacher* concepts are synonyms, so this concepts contract into one node (fig. 4).

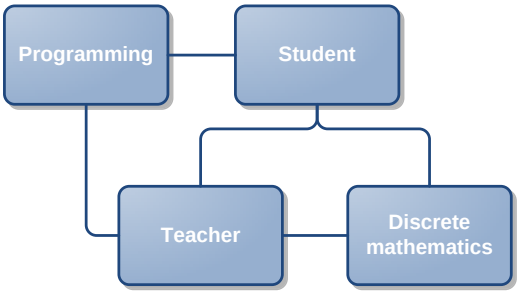

Figure 4. Exam taking semantic network after synonymy reduction

D. Graph coarsening

The next step is to transform the semantic net into ontology model. In general it's graph coarsening problem [5]. Classic methods of solving this problem are based on iterative contraction of adjacent nodes of graph G_α into nodes of graph $G_{\alpha+1}$, where α = 0, 1, 2, … – number of iteration, $G(0) = G(O)$. As a result the edge between two of graph G_α is removed and the multinode of graph $G_{\alpha+1}$ is created. [9].

When two nodes are replaced by one node (during the contraction), the values of these nodes are replaced by the value of parent node from linguistic ontology.

In example *programming* and *discrete mathematics* concepts are coarsened into one node (fig. 5).

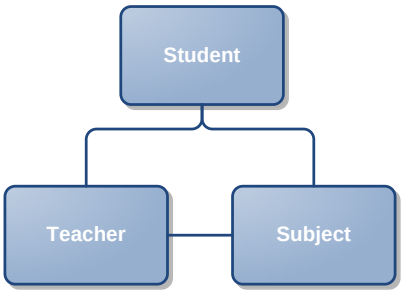

Figure 5. Exam taking semantic network after graph coarsening

E. Ontology improving

At this step we have a base ontology, representing criteria for DSL matching. However, it has some disadvantages:

1) no semantic relations representation;

2) unnecessary concepts may appear (this are useless for current task, but were generated during the analysis);

3) essential concepts could be missed during analysis.

To fix these disadvantages, this base ontology should be edited by human (specify relation semantics, add or delete concepts). Obviously, the more accurate will be ontology model, the more accurate DSL will be matched.

Consider that specialist renamed "*Subject*" to "*Exam*", and removed relation between *student* concept and *teacher* concept, and added the semantic meanings to remaining relations (student takes an exam and teacher grade an exam). The result is shown in fig. 6.

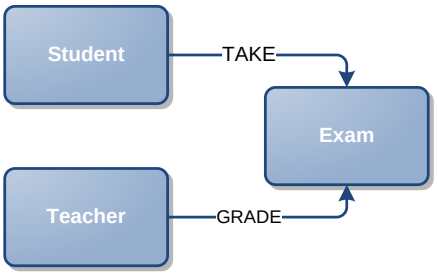

Figure 6. Exam taking ontology

F. Matching evaluation between DSL and created ontology

Comparison of ontologies comes down to calculation or relations revelation between the terms of two ontologies based on different lexical or structure methods. The result of this comparison represents a set of correspondences between the entities that are related semantically.

In order to assess how similar ontologies are, the extent of isomorphism should be measured.

Two graphs $(V1;E1; g1)$ and $(V2;E2; g2)$ are isomorphic if there are bijections:

$$f1 : V1 \to V2 \text{ and } f2 : E1 \to E2$$

so that for each edge

$$\begin{cases} a \in E1 \\ g1(a) = x - y \end{cases}$$

if and only if

$$g2[f2(a)] = f1(x) - f1(y).$$

It is not always easy to establish if two graphs are isomorphic or not. An exception is the case where the graphs are simple. In this case, we just need to check if there is a bijection

$$f: V1 \to V2,$$

which preserves adjacent vertices. If the graphs are not simple, we need more sophisticated methods to check for when two graphs are isomorphic

In our case, we should place emphasis that two graphs are not going to be isomorphic. However, the higher extent of isomorphism is, the more suitable current graph is.

The linguistic ontologies will have huge impact on the extent of isomorphism. For instance, if current node in the first graph was happened to describe a person and current node in the second graph described the document, isomorphism substitution would not exist in this context. At this moment, we are developing linguistic ontology-based algorithm for measuring how isomorphic two graphs are.

IV. Conclusion and future work

In this paper a problem of matching a suitable DSL for specific task was observed.

The requirements for DSL are based on domain documents analysis. Requirements are formed as ontological model which is generated in two steps: defining concepts using frequency analysis of terms found and defining relations based on average weighted score obtained using Jaccard index and mutual information index.

The second step of DSL matching is comparison of DSL's that was implemented earlier with ontology based on domain documents analysis. The core of this comparison is the method of determining graphs' isomorphism and semantic match is controlled by linguistic ontology.

The further work is devoted to increasing the number of methods used to create more relations in the ontology model. This will improve the accuracy of average weighted score of concept relationship. Furthermore the DSL comparison on different levels will be observed (hierarchical structure comparison).

References

[1] A.O. Sukhov, L.N. Lyadova "MetaLanguage: a Tool for Creating Visual Domain-Specific Modeling Languages", Proceedings of the 6th Spring/Summer Young Researchers' Colloquium on Software Engineering, SYRCoSE 2012, Пермь: Институт системного программирования Российской академии наук, 2012, pp. 42-53

[2] Centre for the Analysis of Time Series website. [Online]. Available: http://cats.lse.ac.uk/homepages/liam/st418/mutual-information.pdf

[3] D. Balasubramanian "The Graph Rewriting and Transformation Language: GReAT". [Online]. Available: http://www.isis.vanderbilt.edu/sites/default/files/great_easst.pdf

[4] D. Worth, A. Kozak, D. Johnson "Russian Derivational Dictionary", New York, NY: American Elsevier Publishing Company Inc, 1970

[5] G. Karypis, V. Kumar "Multilevel k-way Partitioning Scheme for Irregular Graphs", Journal of Parallel and Distributed Computing, 96-129, 1998

[6] J. Bezivin "An Introduction to the ATLAS Model Management Architecture". [Online]. Available: http://www.ie.inf.uc3m.es/grupo/docencia/reglada/ASDM/Bezivin05b.pdf

[7] R. Witte, N. Khamis, and J. Rilling, "Flexible Ontology Population from Text: The OwlExporter" Dept. of Comp. Science and Software Eng. Concordia University, Montreal, Canada. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/932_Paper.pdf

[8] R. Real, J. Vargas, "The Probabilistic Basis of Jaccard's Index of Similarity" [Online]. Available: http://sysbio.oxfordjournals.org/content/45/3/380.full.pdf

[9] А. Карпенко "Оценка релевантности документов онтологической базы знаний". [Online]. Available: http://technomag.edu.ru/doc/157379.html

[10] А. Сухов "Методы трансформации визуальных моделей". [Online]. Available: http://www.hse.ru/pubs/share/direct/document/68390345

[11] В. Аверченков, П. Казаков "Управление информацией о предметной области на основе онтологий". [Online]. Available: http://www.pandia.ru/text/77/367/22425.php

[12] В. Гусев "Механизмы обнаружения структурных закономерностей в символьных последовательностях", 47–66, 1983

[13] В. Гусев, Н. Саломатина "Алгоритм выявления устойчивых словосочетаний с учётом их вариативности (морфологической и комбинаторной". [Online]. Available: http://www.dialog-21.ru/Archive/2004/Salomatina.htm

[14] Г. Белоногов, И. Быстров, А. Новоселов и другие "Аавтоматический концептуальный анализ текстов" НТИ, сер. 2, № 10, с. 26–32, 2002

[15] И. Мисуно, Д. Рачковский, С. Слипченко "Векторные и распределенные представления, отражающие меру семантической связи слов". [Online]. Available: http://www.immsp.kiev.ua/publications/articles/2005/2005_3/Misuno_03_2005.pdf

[16] Л. Лядова "Многоуровневые модели и языки DSL как основа создания интеллектуальных CASE-систем". [Online]. Available: http://www.hse.ru/data/2010/03/30/1217475675/Lyadova_LN_2.pdf

[17] М. Гринева, М. Гринев, Д. Лизоркин "Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов". [Online]. Available: http://citforum.ru/database/articles/kw_extraction/2.shtml#3.3

[18] Н. Загоруйко, А. Налётов, А. Соколова и другие "Формирование базы лексических функций и других отношений для онтологии предметной области". [Online]. Available: http://www.dialog-21.ru/Archive/2004/Zagorujko.htmM. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.