International Journal of Systems Engineering
2018; 2(1): 26-28
http://www.sciencepublishinggroup.com/j/ijse
doi: 10.11648/j.ijse.20180201.16

Modeling WordNet Type Thesaurus for Uzbek Language Semantic Dictionary

Matlatipov San'atbek[1], Aripov Mirsaid[1], Abdurakhmonova Nilufar[2]

[1]Applied Mathematics and Computer Analysis Department, National University of Uzbekistan Named After Mirzo Ulugbek, Tashkent City, Uzbekistan

[2]Information and Contemporary Pedagogical Technologies Department, Tashkent State University of Uzbek Language and Literature Named After Alisher Navoiy, Tashkent City, Uzbekistan

Email address:
goodluck_0714@mail.ru (M. San'atbek)

To cite this article:
Matlatipov San'atbek, Aripov Mirsaid, Abdurakhmonova Nilufar. Modeling WordNet Type Thesaurus for Uzbek Language Semantic Dictionary. *International Journal of Systems Engineering*. Vol. 2, No. 1, 2018, pp. 26-28. doi: 10.11648/j.ijse.20180201.16

Received: June 26, 2018; **Accepted**: July 13, 2018; **Published**: August 9, 2018

Abstract: These days creating the corpus of texts for Uzbek language, creating and developing linguistic databases, search-engine systems – are one of the crucial tasks of computational linguistics. Particularly, electronic dictionary-thesauruses, semantic dictionaries are one of them. Dictionary-thesaurus formation structure for Uzbek language, transferring the terminological dictionary into the e-version and implementing rules for establishing semantic relations between words where it gives a chance to establish automation linguistic processes of dictionary-thesauruses, which is the foundation of linguistic databases. Analyzing logical structure of paper-based dictionary thesauruses has given a chance to formalize its structure and creating rules for converting to e-version of dictionary-thesaurus syllables by using predicates language. Descriptors system is suggested in PROLOG language rules set for constructing e-version of dictionary – syllables.

Keywords: Thesaurus, Word, Key Word, Language Sign, Meaning, Database, Prolog Language

1. Introduction

The main aim of the paper is analyzing logical structure of Uzbek dictionary-thesaurus, formation its structure and developing conversion rules of existent paper-based Uzbek dictionaries to electronic dictionary-thesaurus.

Nowadays, constructing Uzbek language thesaurus level semantic dictionaries is not developing in a high temp. One of the main issues in last years as regards NLP activities is the increasingly fast development of generic language resources. A lot of such resources, including both software and lingware items (lexicons, lexical databases, grammars, corpora marked in several ways) have been made available for research and industrial applications [1]. Special interest presents, for knowledge-based NLP tasks, the availability of wide coverage ontologies. Princeton WORDNET, BABLE NET, FRAMENET and European Word NET are considered one of the most known ontologies. The construction of a WordNET for a language depends on the lexical source available. Building the lexical source manually can be very costly. However, it accuracy will be high.

Existent dictionaries are only limited by fulfilling databases [1–3]. However, formalization of Uzbek language linguistics, implementing linguistic processors for automation process of developing electronic thesauruses are still crucial task. Some of the authors for other languages has shown its importance to use XML type special expanded language [5, 6], in turn, this gives a chance to work with structured data. The paper called Building a Wordnet for Turkish [12] which is used the Princeton model to build Turkish wordNET and Uzbek language is also in Turkic language category.

Having structure in abstract lexicographic system is obvious and it has two parts: left (registry) and right (interpretation). Only right sides of the dictionary differentiate its meaning. However, thesaurus has deeper structure where it can establish relations for both (left and right) sides. Thereby, dictionary is such types of text where lexical description of language (s) is described systematic

and structured.

The challenge of building dictionaries is that, it is not always possible to describe exactly all its elements by using above mentioned method. There are very many uncertain elements in the real dictionary structure, which in turn, it will sometimes be challenging problem to solve them.

2. Method

The set of structured elements of dictionary and their association are include the dictionary's meta-language feature. Determining its systematic properties can be a foundation of developing formal model of thesaurus. The process of building dictionary meta-language is the typical method of describing lexicographic.

Info-lexicographic model of any lexicographic system can be described as follows [6]:

$$V(l) = \{\Lambda(l), P(l), H\} \quad (1)$$

Here, $V(l)$ - is the lexicographic system which includes dictionary unit sets; $\Lambda(l)$ - left side part unit set of dictionary; $P(l)$ - right side part unit set of dictionary; H – reflection, that is, fitting the set $\Lambda(l)$ to the set $P(l)$:

$$H: \Lambda(l) \to P(l) \quad (2)$$

H reflection in describing lexicographic system appears by establishing conformity function with its left side to right side and provides dichotomous fullness in building appropriate thesaurus.

Some of the elements is used in the reflection of $H: \Lambda(l) \to P(l)$ in the dictionary and appears in specific terms. The left and right sides of the lexicography are not only formal placement, but also have to implement them with functional relations. Lexicographic sorted, indexed set $S_0(l)$ is determined in $V(l)$ set.

We can observe that, current Uzbek language dictionaries are $S_0(l)$ one element set. Resemble to paper-based dictionaries, lexicographic model structure also have same approach that begins from L headword and it serves as an identification (ID) of dictionary unit in lexicographic system.

The reasons of strong agglutinative existence in building Uzbek language, words are constructed by adding prefixes and suffixes. For that reason, it is important to mention these relations separately for thesaurus dictionary, as an example, "mansab"-"mansabdor", "suv-suvchi".

$$D: \Lambda(l) \to V(l) \quad (3)$$

As we know, Princeton WordNET [7] system has five types of syntactic categories: They are NOUN, VERB, ADJECTIVE, ADJECTIVE SATELLITE, and ADVERB. Accordingly, building WordNET type thesaurus for Uzbek language, we add signs for syntactic categories of the following part of speeches: Ot (Noun), Sifat (Adjective), Son (Number), Fe'l (Verb), Ravish (Adverb), Olmosh (Pronoun), Undov (Exclamation), Modal (Modal verbs), Taqlid (imitation), Yuklama (Particle), Ko'makchi (accessorial), Bog'lovchi (Conjunction). As one of constructing elements of main system of $V(l)$ lexicographic system, we can show its automorphism system, that is, reflection to itself of $V(l)$ system:

$$A: \Lambda(l) \to V(l) \quad (4)$$

3. Result

We can determine type of pointer of A automorphism dictionary syllables, for example, "$x\ syn.y$". This type automorphism of dictionary syllables determines such reflection, where it is as follows:$\mathrm{V}(x) \to V(y)$. As its ID, usually any of pointer pseudo-word is used (in the example – "$syn.y$"), thus, it correspondingly puts $V(y)$ to $\mathrm{V}(x)$. Need to mention that, constructing A automorphism is more complex than above example. Firstly, number of pointer can be more than one, that is to say, it can have recurrent property:

Moreover, automorphic reflection can be constructed as following:

$$V(x) \to \{V(x')\} \to \cdots \to \{V(x''')\} \to \cdots \quad (5)$$

$$x, x', x'', \ldots sm.y, y', y'', \ldots. \quad (6)$$

In Uzbek language also, thesaurus structure elemental unit is – dictionary unit of descriptor and it is constructed as alphabetical order. We can describe thesaurus dictionary unit as following for Uzbek language:

$$d_i < M_{i1}, M_{i2}, M_{i3},\ M_{i4}, M_{i5} > \quad (7)$$

Here d_i the title descriptor; M_{i1} – is alphabetically sorted conditional synonyms set of given title descriptor and they together consist conditional equivalence class; M_{i2} – in every title descriptor is connected with "tur-mansub" relation (hyperonym, hyponym) and alphabetically sorted descriptors set; M_{i3} – in every title descriptor is connected with "butun – qism" relation (member, member-of, meronym) and alphabetically sorted descriptors set; M_{i4} – in every title descriptor is connected with at least following paradigmatic relation and alphabetically sorted descriptors set: "sabab – oqibat", "oqibat – sabab", "funksional moslik" (associative relations); M_{i5} – in every title descriptor is connected with "antonim"(antonym) relation and alphabetically sorted descriptors set; M_{i6} – because of strong agglunativity in every title descriptor is connected with affixes (constructing words by adding affixes) and alphabetically sorted descriptors set;

These relations establish relationships of X word to Y word. By these relationships, semantic net of language is constructed. Any one of presented sets can be single value or empty, dictionary even may not be in unit.

4. Discussion

The set M_{i1} consist conditional equivalence class together with d_i title description and it is also descriptor. This M_{i1} set

plays a role in nominal definition function and makes meaning determine of d_i title descriptor meaning in conditional equivalence class.

Taking into account the fact that most of the information base of the Internet is based on symbolic information, it is desirable to use Prolog [8] to use the logical programming language to work with symbolic structures, text files, and intelligent computer programs. Prolog is an easy programming language to describe objects and their relationships when looking for a solution.

By considering dictionary unit structure [7] of dictionary-thesaurus, predicate-rules are implemented in PROLOG programming for every dictionary unit:

$sinset(X_{i1}, X_{i2})$ predicate is put to M_{i1} synonyms set correspondingly;

$hyp(X_{i1}, X_{i2})$ predicate is put to M_{i2} "tur – mansub" relation among its relations; $mer(X_{i1}, X_{i2})$ predicate is put to M_{i3} "butun – qisim" relation among its relations; $cause(X_{i1}, X_{i2})$ predicate is put to M_{i4} "sabab - oqibat" relation among its relations; $ant(X_{i1}, X_{i2})$ predicate is put to M_{i5} antonym relation among its relations; $der\ (X_{i1}, X_{i2}, A)$ predicate is put to M_{i6} which is the function of constructing new words from existing ones. As a result, we can gain semantic set of dictionary unit by implementing predicate-rules for PROLOG [12-13].

5. Conclusion

Analyzing logical structure of paper-based dictionary thesauruses has given a chance to formalize its structure and creating rules for converting to e-version of dictionary-thesaurus syllables by using predicates language. Descriptors system is suggested in PROLOG language rules set for constructing e-version of dictionary – syllables.

The model of prepared dictionary-thesaurus and the rules for converting dictionary-thesaurus into e-version form can be a base of constructing lexicographic process. When forming a lexicographical database, the elements of the system structure are defined as the database elements and its search parameters. The process of forming the lexicographical database on the basis of forming the system elements of the system leads to a fully automated procedure.

The dictionary can be used as part of the linguistic support of an automated system built into a suitable subject area.

References

[1] Gerd A. S. Databases and applied linguistics. Reports of the scientific conference "Corpus linguistics and linguistic databases" / Pod. Ed. A. S. Gerda. - St. Petersburg.: Publishing house S.-Petersburg. University, 2002. - 168 p.

[2] Azarova I. V. and others. Development of a computer thesaurus of the Russian language such as WordNet. Reports of the scientific conference "Corpus linguistics and linguistic databases" / Pod. Ed. A. S. Gerda. - St. Petersburg.: Publishing house S.-Petersburg. University, 2002. - 168 p.

[3] Azarova L., and others. RussNet: Building a Lexical Database for the Russian Language // Proceedings of Workshop on WordNet Applications and Evaluation in LREC2002, June 2002. Las Palmas de Gran Canaria, 2002.

[4] O. Bilgin, O. Cetinoglu, K. Oflazer. "Building A WordNet For Turkish". Romanian Journal Of Informatıon Science And Technology, Volume 7, Numbers 1-2, 2004, 163-172.

[5] Kasilov O. V. Methods for presenting structured texts of natural language in XML description // News of NTU "KhPI". Zbirka naukovyh prat. Thematic schedule: Novi rishennya u perchasnih tehnologiïy. - Khar'k: NTU "KhPI". - 2002. - No. 6. - T. 2 - 156 p.

[6] Kasilov O. V. Modeling of the dictionary-thesaurus // Vestnik NTU KhPI. 2004. №34.

[7] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[8] Michael A. Covington, Prentice Hall, 1994 Natural Language Processing for Prolog Programmers.

[9] Matlatipov G., Vetulani Z. (2009) Representation of Uzbek Morphology in Prolog. In: Marciniak M., Mykowiecka A. (eds) Aspects of Natural Language Processing. Lecture Notes in Computer Science, vol 5070. Springer, Berlin, Heidelberg.

[10] Güngör, T., Kuru, S.: Representation of Turkish morphology in ATN. Booaziçi University, Istanbul (1993).

[11] Güngör, T., Kuru, S.: Representation of Turkish morphology in ATN. Booaziçi University, Istanbul (1993).

[12] Xavier Ferreres, German Rigaue, Horacia Rodriguez, Using WordNET or building WordNETs. p 65-72.

[13] Orhan Bilgin, Ozlem CETINOGLU, Kemal OFLAZER, Building a Wordnet for Turkish, Romanian journal of Information science and technology, Vol 7, 2004, 163-172.