Intelligence Knowledge-basedSystem Based on Multilingual Dictionaries

Oleksii Puzik[0000-0003-1460-1686]

Kharkiv National University of Radio Electronics
14, Naukyave., 61000, Kharkiv, Ukraine

alphabet308@gmail.com

Abstract. Intelligence knowledge-based systems are important part of natural language processing researches. Appropriate formal models simplify developing of such systems and open new ways to improve their quality. This work is devoted to developing of intelligence knowledge-based system using model based on algebra of finite predicates. The model also isbased on lexicographical computer system which consists of trilingual and explanatory dictionaries. Algebra of finite predicates is used as formalization tool.Problems of distinguishing semantic entities is investigated during research. Method of resolving homonymy ambiguities is used to extract separate entities, thus allowing formalization of semantic relationships. In result formal model of intelligence knowledge-based system was developed.It was shown way to extend the model for different languages.

Keywords: Natural Language Processing, Intelligence Knowledge-based Systems, Algebra of Finite Predicates, AFP, Lexicographical System

1 Introduction

Natural language modeling is important partof the theory of intelligence. The main objects of the language description are words and relations between them.The question is understanding texts written in natural language, create mathematical models for solving linguistic problems and developing programs that function based on these models. The construction of intelligence knowledge-based systems requires the automation of both the syntax of the natural language and its semantics. This part of computational linguistics relates to the section of artificial intelligence, engaged in the development of natural language word processing systems.This problem despite the different approaches to the formalization of semantic problems is still not completely solved because of lack of formalization while describing semantic relations.

This paper proposes formal description of a data model based on algebra of finite predicates (AFP) [1].Thesystem can be used as a part of intelligence system which is based on trilingual terminological dictionary on computer science and radio electronics.

COLINS’2019, Volume II: Workshop. Kharkiv, Ukraine, April 18-19, 2019, ISSN 2523-4013 http://colins.in.ua, online

Also, in the study considered formalization of the problem of homonymy ambiguation. There are exist many different approaches [2, 3] regarding multiple-valued structures. Suggested approach allows to resolve homonymy ambiguation by adding explanatory dictionaries to semantic concepts and calculating relevance score based on those explanations.

2 Trilingual dictionary model

Let's assume model of trilingual dictionary as a set of words in different languages for certain knowledge area. In this study we use Ukrainian-Russian-English dictionary of radio electronics and informatics. We need to group all these words by word meanings to build intelligence knowledge-based system. In general case it is hard to achieve this because we have homonymy ambiguity. The classic example is words "*spring*"in English or "*коса*" in Russian and Ukrainianwhich have few meanings. Thus, if we build intelligence system based only on spelling of these words, we can't distinguish semantic. Moreover, if need to get translations we face with issue that invalid translation equivalents pair can be matched. Let's introduce additional semantic metalanguage which has exactly one definition for each semantic conception.In this case we can groupwords from different languages using these abstract words from the semantic metalanguage. Graphical representation is displayed in Fig. 1.

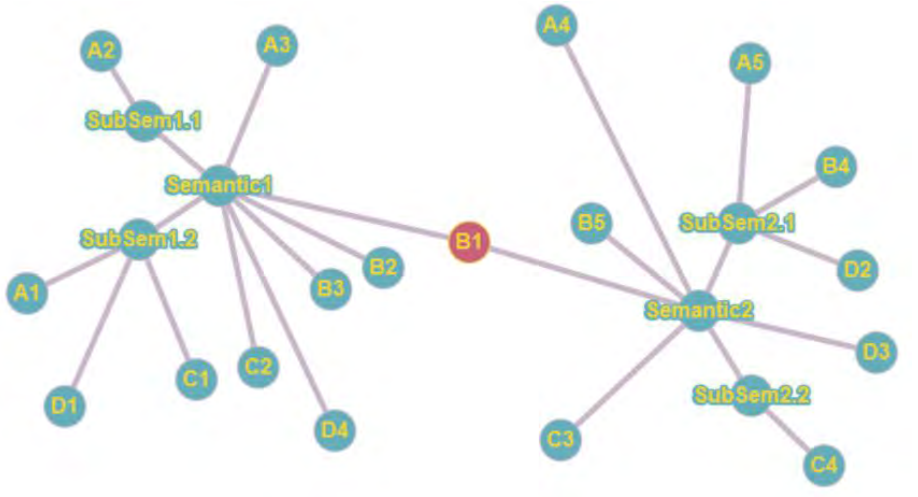


Fig. 3. Graphical representation of words in multilingual dictionary with metalanguage

In this figure Ai, Bi, Ci, Di are some words from different languages where letter identifies some language and index corresponds to i-th word in the language introduced in the dictionary, Semantici – is a semantic concept which connects corresponding words, Semantici, j – is a subsemantic concept which emphasizes semantic features specific to some language.

Thus, we create additional level of indirection. This way used widely in software engineering. For instance, it can be used for machine translations where no direct dictionary between languages but exist two or more dictionaries with common language used in both [4].

Additional feature of such metalanguage is that it is possible to attach explanatory dictionary to the metalanguage entity node and it will have just only one the most appropriate explanation per language per metaword. Also, we can introduce additional features like subsemantics or shadows of meaning attached to main semantic thus we can introduce more specific translations for each language.

Let D – set that represents some dictionary, W – set of all words for all languages, S – set of semantic concepts.

$$D = S_1 \ \vee S_2 \ \vee \ldots \vee S_n$$

Let w$_{Si}$ ⊆ W – subset of words that represents certain semantic conception S$_i$

$$S_i = w_i^{sem} \vee w_i^{ru} \vee w_{i1}^{ru} \vee w_i^{ua} \vee w_i^{en} \ ,$$

where w$_i$ – words related for certain semantic conception S$_i$for Semantic metalanguage, Russian, Ukrainian and English accordingly. Also, these words may include synonymy words like w_{i1}^{ru} because despite different spelling they represent the same semantic concept.

The following predicate evaluates the equivalence for any two words in the dictionary:

$$P_e(x \ \ , y \ \ ) = \begin{cases} 1, \ (x \ \ , y \ \ ) \in S_i \\ 0, \ (x \ \ , y \ \ ) \notin S_i \end{cases}$$

At the same time $w_i^{sem} \in S_i$ by its definition. So, we can write predicate which introduces translations and synonymy in our dictionary:

$$P_e(x \ \ , y \ \ ) = P_e(x \ \ , w_i^{sem}) \wedge P_e(y, w_i^{sem})$$

Equations written above have one issue – they should work on sets of words without intersections, but natural languages mostly have some homonymouswords. Some such word is shown as B1 node in the Fig. 1. Thus, we need to find out method to separate homonymous words. So, in result we need to get following:

$$S_{spring1} = w_{spring1}^{sem} \vee w_{\text{пружина}}^{ru} \vee w_{\text{пружина}}^{ua} \vee w_{spring1}^{en}$$

$$S_{spring2} = w_{spring2}^{sem} \vee w_{\text{весна}}^{ru} \vee w_{\text{весна}}^{ua} \vee w_{spring2}^{en}$$

However, in general case if we just look for translation of separate word, we can get list of possible semantic conceptsof that word. The list can be used to get translations specific for certain language.

3 Homonymy disambiguation

In real world texts we don't have such specific marks like spring1. Moreover, even they exist they will not match our marks in our dictionary. Above we mentioned that semantic nodes may be attached with explanatory nodes. In this case we may consider these explanatory nodes as the same logical entities as word nodes.

$$S_i = w_i^{sem} \vee w_i^{ru} \vee w_{i1}^{ru} \vee w_i^{ua} \vee w_i^{en} \vee x_i^{ru} \vee x_i^{ua} \vee x_i^{en},$$

where w_i – words related for certain semantic conception S_i for Semantic metalanguage, Russian, Ukrainian and English accordingly,x_i – explanations related to certain semantic conception S_i for Russian, Ukrainian and English accordingly.Also, these explanation nodes will help us calculate relevance score for homonymous words.

Homonymy disambiguation is not possible without considering context where homonymous word is used.One of principles which must be used when defining explanations for each sematic concept is that all word used in explanation must be present in the dictionary [5]. We can useapproach described in [6] to select homonym the most relevantto context. In few words, relevance score is calculated for each word used in the context regarding processed word.

Let define context function C(*x, ctx*) which calculatesrelevancescore, where *x* – is some word and *ctx* – is the context where the word *x*is used. Let E(*x*) function which calculates relevance score based on explanations used in the dictionary.Thus we can define function

$$\mathrm{G_i}(x_i, y, ctx) = |\mathrm{E}(x_i) - \mathrm{C}(y, ctx)|,$$

where x_i– word with potential semantic connected to*y*, *y* – word which semantic should be determined, *ctx*–context where word *y* is used.

$$\mathrm{F}(y, ctx) = \min_{i=1,n} \mathrm{G}_i(x_i, y, ctx)$$

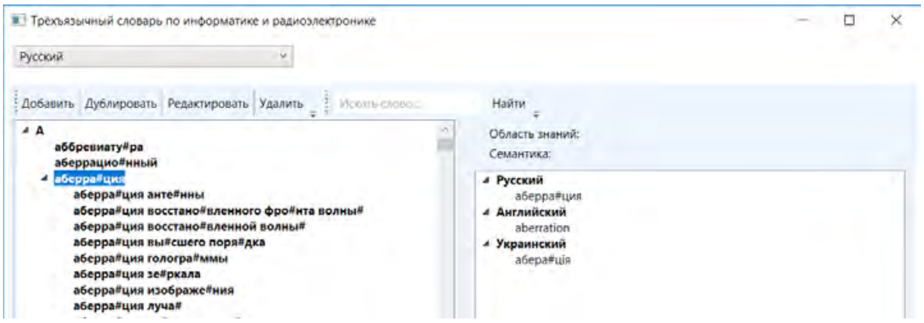
where$\mathrm{F}(y, ctx)$is a function which calculatesthe most relevant semantic concept for word *y* based on context where the word is used.

Cognate languages may have similar words but with different semantic. The suggested method may be extended by including explanatory dictionaries for different languages into intelligence knowledge-based system.This allows avoiding such kind of ambiguities.

4 Trilingual electronic dictionary

Based on suggested approach electronic trilingual dictionary of informatics and radioelectronics was developed. The dictionary can be usedby the intelligence knowledge-based systemas a semantic database.

The software system of the trilingual electronic dictionary is based on the paper version of the Russian-Ukrainian terminological dictionary on informatics and radio electronics. An electronic version of the text was obtained by scanning and recognition, which serves as the basis for filling the dictionary with content. The electronic version of the dictionary allows user to add translation equivalents of terms for the English language and proceed to the multilingual dictionary on further development.



Трёхъязычный словарь по информатике и радиоэлектронике

Русский

- A
  - аббревиату#ра
  - аберрацио#нный
  - аберра#ция
    - аберра#ция анте#нны
    - аберра#ция восстано#вленного фро#нта волны#
    - аберра#ция восстано#вленной волны#
    - аберра#ция вы#сшего поря#дка
    - аберра#ция гологра#ммы
    - аберра#ция зе#ркала
    - аберра#ция изображе#ния
    - аберра#ция луча#

Область знаний:

Семантика:

- Русский
  - аберра#ция
- Английский
  - aberration
- Украинский
  - абера#ція

Fig. 4. Trilingual dictionary UI

The software system of the trilingual dictionary is written in the C# programming language. The architecture is built using the MVVM design pattern (Model-View-ViewModel Model-View-View Model). Thus the software system is clearly divided into separate independent modules. This approach allows reusing and independent changing of individual components of the software system. The user interface is created using WPF technology [7].

The software system of the trilingual electronic dictionary allows user to view, search, edit, add, delete concepts of terms and their translation equivalents for each of languages. The advantage of the proposed approach is the possibility of free switching between the languages of the dictionary and quick access to all translation equivalents of the selected term which corresponds to certain conception semantic.

5 Conclusions and future work

In results was created software system of trilingual terminological Ukrainian-Russian-English dictionary of radio electronics and informatics. The developed model allows quick recognition of synonymous words and translation equivalents. In the study was considered homonymy ambiguities problem and suggested ways to resolve the issue. Suggested model allows extending intelligence system with new languages by not only adding new terms but explanations and specific semantic concepts as well. Intelligence knowledge-based system will use this developed electronic dictionary as a database of low-level terms.

Further development will integrate explanatory dictionary to the intelligence system. Thus, context for each word will be defined and this allows matching context from arbitrary text with semantic concepts stored in dictionary based on suggested approach for homonymy disambiguation.

Long-term future work should consider includingmodules related to processing blocks of texts with extracting complex semantic entitieswhich represent more abstract concepts.

References

1. Bondarenko, M., Shabanov-Kushnarenko, Y.:Teoriyaintellekta: ucheb. Izd-vo SMIT, Kharkov (2006).
2. Chetverikov, G., Vechirska, I.,Tanyanskiy, S.: The methods of algebra of finite predicates in the intellectual system of complex calculations of telecommunication companies. In:2014 24th International Crimean Conference Microwave & Telecommunication Technology, pp. 346-347, Sevastopol (2014).
3. Chetverikov, G., Puzik, O.,Vechirska, I.: Multiple-valued structures of intellectual systems. In: 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), pp. 204-207, Lviv (2016).
4. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F.B., Wattenberg, M., Corrado, G.S., Hughes, M., Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation., Transactions of the Association for Computational Linguistics, vol. 5, 339-351, https://www.transacl.org/ojs/index.php/tacl/article/view/1081, last accessed 2019/03/20.
5. Shirokov, V., Computer lexicography. Scientific and publishing enterprise "Vidavnitstvo"Naukovadumka" NAN Ukrainy",Kyiv(2011).
6. Chetverikov G., Puzik, O.,Tyshchenko,O.: Analysis of the Problem of Homonyms in the Hyperchains Construction for Lexical Units of Natural Language. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), pp. 356-359, Lviv (2018).
7. Chetverikov, G., Vechirska, I., Puzik, O.:Technical Features of the Architecture of an Electronic Trilingual Dictionary. Cognitive studies (16), 143-152 (2016).