Modelling of the Context Links Between the Natural Language Sentences

Anastasiia Vavilenkova [0000-0002-9630-4951]

National aviation university, Kyiv, Ukraine
vavilenkovaa@gmail.com

Abstract. In this materials author proposed the rules for searching the context links between the natural language sentences. This theory used mathematical apparatus for the formal description of an electronic text document based on the predicate logic, which, unlike formal grammars, makes it possible to structure textual information, starting with the lowest level of constructing logical relationships and ending with the text as a whole. These rules help to find context links according to the semantic reiteration, that is used, when text in natural language sentence is formed. Author tried to solve a part of the problem of extraction knowledge from the textual information. The study demonstrates different examples of semantic reiteration usage: tautological reiteration, thematic reiteration and reiteration of various stylistic interpretations. Depending on this example, there are following replacements in logic and linguistics models: relation into relation, subject into subject, object or matter-subject and object into subject, object or matter-subject.

Keywords: natural language, logic and linguistic models, text information, content modelling, context links

1 Introduction

Nowadays computer linguistic is one of the most essential tool for solving the problem of knowledge extraction from the textual information. This mechanism can integrate computer modelling, mathematical methods and linguistic rules [1,3,4]. Machine learning, data science and natural language processing, like the most popular spheres of knowledge extraction are widely spread in different social human areas. According to the IBM predicts, 59% of all data science and analytics job demand is in finance and insurance, professional services, and IT. Data science and analytics job market in Germany, in 2019 needed 15% of seniors. More than 60% of job openings require middle-level specialists, while around quarter offers look for seniors into the job market in Switzerland.

The world is seeing a surge in demand for data science services in various field with market researches estimating its potential growth in the near future. In India, 70% of job postings in this sector are for data scientists with less than five years of work

experience. So, data science is very popular nowadays, there are great number of program products that try to model semantic links. For example, program ABBYY Compreno, which based on the logical derivation and syntactical rules of natural language sentences building [2].

Adam Geitgey shows in his article “Natural language processing is fun” [5] how computer can understand human language. All steps of this process, such as sentence segmentation, word tokenization, predicting parts of speech for each token, text lemmatization, identifying stop word, dependency parsing, finding noun phrases, named entity recognition and co-reference resolution, have detail description.

Real experience has been written in the article “How I used natural language processing to extract context from news headlines” by Gunnvant Saini [8], who tried to extract interesting information from a large number of news documents.

In the article “Affected Experiencers” [10] authors proposed formal analysis of the affected experiencer construction, illustrated by the following examples in german, albanian, japanese and hebrew and tried to recognize semantic of the natural language sentences.

Many Ukrainian scientists develop areas of data science in different ways. For instance, academic of NAS of Ukraine Shyrokov V.A. created linguistic corpus for ukrainian natural language [9], Glybovets M. M. solves the problem by applying various genetic algorithms [6], Lande D.V. used networks and respondents’ perception for searching data [7] etc.

However, all these facts don’t work out with searching a context between different parts of textual electronic document. Unsolved part of the problem is how to find instruments and methods for modelling and extraction knowledge from the textual information.

The aim of the research is formulation and modelling rules for searching the context links between the natural language sentences by means of logic and linguistic models [12].

2 Materials and Methods

It is said, that every simple sentence of natural language, that consist indivisible content, can be presented as a general logic and linguistic model [11]:

$$L_p^S = p(x, g, y, q, z, r, h), \tag{1}$$

where S is natural language sentence;

– p - predicate, that indicates content of the sentence, relation, that connect subject, object and subject-matter of relations in the sentence S ;

– x - subject of the sentence S ;

– g - characteristic of the subject of the sentence S ;

– y - object of the sentence S ;

– q - characteristic of the object of the sentence S ;

– z - subject-matter of the sentence S ;

– r - characteristic of the subject-matter of the sentence S;
– h - characteristic of the p relation in the sentence S.

Naturally, that complex sentence might have been combined different combinations of logic and linguistic models (1). So, if we have textual fragment, that includes two sentences of natural language, and they can be represented by means of decomposition of formal models (1):

$$L^{S_1} = \bigwedge_{\lambda=1}^{\upsilon^{S_1}} L^{S_1}_{p^{(\lambda)}}(h), \tag{2}$$

$$L^{S_2} = \bigwedge_{\mu=1}^{\upsilon^{S_2}} L^{S_2}_{p^{(\mu)}}(h), \tag{3}$$

where

- $L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)})$ – simple predicate, which describes the part of the sentence S_1, that reflects indivisible content;
- $L^{S_2}_{p^{(\mu)}}(h) = p_2^{(\mu)}(x_2^{(\mu)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, z_2^{(\mu)}, r_2^{(\mu)}, h_2^{(\mu)})$ – simple predicate, which describes the part of the sentence S_2, that reflects indivisible content;
- $\lambda = \overline{1, \upsilon^{S_1}}$, υ^{S_1} – the amount of parts in the sentence S_1, that have indivisible content;

$\mu = \overline{1, \upsilon^{S_2}}$, υ^{S_2} – the amount of parts in the sentence S_2, that have indivisible content.

Searching for context linkers in textual fragment means filling the array of additional characteristics for each simple natural language sentence. This materials offer to find context linkers between two complex sentences S_1 and S_2 according to the rules of the content formation in inflecting natural languages, that are called as semantic reiteration. Semantic reiteration is the technique for location sentences with similar meaning. Syntax structure of these sentences will be the same, but relation, subject or object may be replaced by synonyms. It is possible to implement semantic reiteration in several different ways. In all situations the order of the sentences does not matter. The simplest type of the semantic reiteration is **tautological reiteration** – the elementary arrangement of link by using of identity words or word forms, having one word root. If relations in two sentences of natural language $p_1^{(\lambda)} \equiv p_2^{(\mu)}$ or $\hat{p}_1^{(\lambda)} \equiv \hat{p}_2^{(\mu)}$, where $\hat{p}_1^{(\lambda)} \in \hat{R}_p$ and $\hat{p}_2^{(\mu)} \in \hat{R}_p$, that relate to the same range of words with identity radical $\hat{R}_p \in R$ from the set R, or $p_1^{(\lambda)} \in W_p$ and $p_2^{(\mu)} \in W_p$, that relate to the same synonymic range $W_p \subseteq W$ from the set W, so it is necessary to make replacement of $p_1^{(\lambda)} = p_2^{(\mu)}$ into the logic and linguistic models L^{S_1} and L^{S_2} after what:

$$L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$

$$L^{S_2}_{p^{(\mu)}}(h) = p_1^{(\lambda)}(x_2^{(\mu)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, z_2^{(\mu)}, r_2^{(\mu)}, h_2^{(\mu)}).$$

In this regard the elements of arrays of characteristics for both sentences will be: $l_i^{S_1} = p_1^{(\lambda)}$, $i = \overline{1, N_1}$ and $l_j^{S_2} = p_1^{(\lambda)}$, $j = \overline{1, N_2}$, where N_1, N_2 – amount of all elements of arrays of characteristics in the natural language sentences S_1 and S_2.

For instance, logic and linguistic model for textual fragment “Indians who settled in northern areas hunted and fished. Those who settled in the east and southwest farmed” will be:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)}, q_1^{(1)},0,0,0) \,\&\, p_1^{(2)}(x_1^{(2)},0,0,0,0,0,0) \,\&\, p_1^{(3)}(x_1^{(3)},0,0,0,0,0,0).$$

L^{S_1} = settled (indians, 0, areas, nothern, 0, 0, 0)
& hunted (who, 0, 0, 0, 0, 0, 0) & fished (who, 0, 0, 0, 0, 0, 0).

$$L^{S_2} = p_2^{(1)}(x_2^{(1)},0, y_2^{(1)},0,0,0,0) \,\&\, p_2^{(2)}(x_2^{(2)},0, y_2^{(2)},0,0,0,0) \,\&\, p_2^{(3)}(x_2^{(3)},0,0,0,0,0,0).$$

L^{S_2} = settled (who, 0, east, 0, 0, 0, 0) & settled (who, 0, southwest, 0, 0, 0, 0) &
farmed (who, 0, 0, 0, 0, 0, 0).

According to the identity conditions of logic and linguistic models [12] it is possible to do such replacement: $x_1^{(2)} = x_1^{(1)}$, $x_1^{(3)} = x_1^{(1)}$, $y_1^{(2)} = y_1^{(1)}$, $y_1^{(3)} = y_1^{(1)}$, $q_1^{(2)} = q_1^{(1)}$, $q_1^{(3)} = q_1^{(1)}$, $g_1^{(2)} = p_1^{(1)}$, $g_1^{(3)} = p_1^{(1)}$, $x_2^{(1)} = p_2^{(1)}$, $x_2^{(3)} = p_2^{(1)}$, $x_2^{(4)} = p_2^{(1)}$, $x_2^{(2)} = x_2^{(1)}$, $x_2^{(3)} = x_2^{(1)}$, $x_2^{(4)} = x_2^{(1)}$, $y_2^{(3)} = y_2^{(1)}$, $y_2^{(4)} = y_2^{(2)}$, $g_2^{(3)} = p_2^{(1)}$, $g_2^{(4)} = p_2^{(1)}$.

After transforming logic and linguistic models will be:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)}, q_1^{(1)},0,0,0) \,\&\, p_1^{(2)}(x_1^{(1)}, p_1^{(1)}, y_1^{(1)}, q_1^{(1)},0,0,0) \,\&$$
$$p_1^{(3)}(x_1^{(1)}, p_1^{(1)}, y_1^{(1)}, q_1^{(1)},0,0,0).$$

L^{S_1} = settled (indians, 0, areas, nothern, 0, 0, 0)
& hunted (indians, settled, areas, nothern, 0, 0, 0) &
fished (indians, settled, areas, nothern, 0, 0, 0).

$$L^{S_2} = p_2^{(1)}(x_2^{(1)},0, y_2^{(1)},0,0,0,0) \,\&\, p_2^{(2)}(x_2^{(1)},0, y_2^{(2)},0,0,0,0) \,\&$$
$$p_2^{(3)}(x_2^{(1)}, p_2^{(1)}, y_2^{(1)},0,0,0,0) \,\&\, p_2^{(3)}(x_2^{(1)}, p_2^{(1)}, y_2^{(2)},0,0,0,0).$$

L^{S_2} = settled (who, 0, east, 0, 0, 0, 0) & settled (who, 0, southwest, 0, 0, 0, 0) &
farmed (who, settled, east, 0, 0, 0,0) & farmed (who, settled, southwest, 0,0,0,0).

According to the rule:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)}, q_1^{(1)},0,0,0) \,\&\, p_1^{(2)}(x_1^{(1)}, p_1^{(1)}, y_1^{(1)}, q_1^{(1)},0,0,0) \,\&$$
$$p_1^{(3)}(x_1^{(1)}, p_1^{(1)}, y_1^{(1)}, q_1^{(1)},0,0,0).$$

$$L^{S_2} = p_1^{(1)}(x_1^{(1)},0, y_2^{(1)},0,0,0,0) \,\&\, p_1^{(1)}(x_1^{(1)},0, y_2^{(2)},0,0,0,0) \,\&$$
$$p_2^{(3)}(x_1^{(1)}, p_1^{(1)}, y_2^{(1)},0,0,0,0) \,\&\, p_2^{(3)}(x_1^{(1)}, p_1^{(1)}, y_2^{(2)},0,0,0,0).$$

L^{S_1} = settled (indians, 0, areas, nothern, 0, 0, 0)

& hunted (indians, settled, areas, nothern, 0, 0, 0) &
fished (indians, settled, areas, nothern, 0, 0, 0).

L^{S_2} = settled (indians, 0, east, 0, 0,0,0) &
settled (indians, 0, southwest, 0,0,0,0) &
farmed (indians, settled, east, 0,0,0,0) &
farmed (indians, settled, southwest, 0,0,0,0).

The result of the rule applying is: $l_i^{S_1} = p_1^{(1)}$, $l_{i+1}^{S_1} = z_1^{(1)}$ and $l_j^{S_2} = p_1^{(1)}$, $l_{j+1}^{S_2} = z_1^{(1)}$. And according to the next rule: $l_{i+2}^{S_1} = x_1^{(1)}$ and $l_{j+2}^{S_2} = x_1^{(1)}$.

Another type of the semantic reiteration thematic reiteration - the words in the sentences show common lexical meaning, indicate different sizes, components, parts of the elements from one situation. The word inside of the one thematic group make the paradigm, which connect various parts of the text.

If subjects of the relations in two natural language sentences are similar $x_1^{(\lambda)} \equiv x_2^{(\mu)}$ or $\hat{x}_1^{(\lambda)} \equiv \hat{x}_2^{(\mu)}$, where $\hat{x}_1^{(\lambda)} \in \hat{R}_x$ and $\hat{x}_2^{(\mu)} \in \hat{R}_x$, $\hat{R}_x \in R$, or $x_1^{(\lambda)} \in W_x$ and $x_2^{(\mu)} \in W_x$, that relate to the same synonymic range $W_x \subseteq W$ from the set W, so it is necessary to make replacement of $x_1^{(\lambda)} = x_2^{(\mu)}$ into the logic and linguistic models L^{S_1} and L^{S_2} after what:

$$L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$
$$L^{S_2}_{p^{(\mu)}}(h) = p_2^{(\mu)}(x_1^{(\lambda)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, z_2^{(\mu)}, r_2^{(\mu)}, h_2^{(\mu)}).$$

The elements of arrays of characteristics for both sentences will be: $l_i^{S_1} = x_1^{(\lambda)}$, $i = \overline{1, N_1}$ and $l_j^{S_2} = x_1^{(\lambda)}$, $j = \overline{1, N_2}$, where N_1, N_2 – amount of all elements of arrays of characteristics in the natural language sentences S_1 and S_2.

If subject, object or matter-subject from any one sentence is the same as the matter-subject of another sentence of natural language, that means they are identical $\hat{x}_1^{(\lambda)} \equiv \hat{z}_2^{(\mu)}$ or $\hat{y}_1^{(\lambda)} \equiv \hat{z}_2^{(\mu)}$, or $\hat{z}_1^{(\lambda)} \equiv \hat{z}_2^{(\mu)}$, or $\hat{x}_1^{(\lambda)} \equiv \hat{x}_2^{(\mu)}$, or $\hat{y}_1^{(\lambda)} \equiv \hat{x}_2^{(\mu)}$, or $\hat{z}_1^{(\lambda)} \equiv \hat{x}_2^{(\mu)}$, where $\hat{x}_1^{(\lambda)} \in \hat{R}_x$, $\hat{y}_1^{(\lambda)} \in \hat{R}_x$, $\hat{z}_1^{(\lambda)} \in \hat{R}_x$, $\hat{x}_2^{(\mu)} \in \hat{R}_x$ and $\hat{z}_2^{(\mu)} \in \hat{R}_x$, that relate to the same range of words with identical root $\hat{R}_x \in R$ from the set R, or $x_1^{(\lambda)} \in W_x$, $y_1^{(\lambda)} \in W_x$, $z_1^{(\lambda)} \in W_x$, $x_2^{(\mu)} \in W_x$ and $z_2^{(\mu)} \in W_x$, that relate to the same synonymic range $W_x \subseteq W$ from the set W, so it is necessary to make replacement of $x_2^{(\mu)} = x_1^{(\lambda)}$ or $x_2^{(\mu)} = y_1^{(\lambda)}$, or $x_2^{(\mu)} = z_1^{(\lambda)}$, or $z_2^{(\mu)} = x_1^{(\lambda)}$ or $z_2^{(\mu)} = y_1^{(\lambda)}$, or $z_2^{(\mu)} = z_1^{(\lambda)}$ into the logic and linguistic models L^{S_1} and L^{S_2} after what:

$$L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$

$$L^{S_2}_{p^{(\mu)}}(h) = p_2^{(\mu)}(x_1^{(\lambda)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, x_1^{(\lambda)}, r_2^{(\mu)}, h_2^{(\mu)})$$

or

$$L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$

$$L^{S_2}_{p^{(\mu)}}(h) = p_2^{(\mu)}(x_1^{(\lambda)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, y_1^{(\lambda)}, r_2^{(\mu)}, h_2^{(\mu)}),$$

or

$$L^{S_1}_{p^{(\lambda)}}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$

$$L^{S_2}_{p^{(\mu)}}(h) = p_2^{(\mu)}(x_1^{(\lambda)}, g_2^{(\mu)}, y_2^{(\mu)}, q_2^{(\mu)}, z_1^{(\lambda)}, r_2^{(\mu)}, h_2^{(\mu)}).$$

The elements of arrays of characteristics for both sentences will be: $l_i^{S_1} = x_1^{(\lambda)}$ and $l_j^{S_2} = x_1^{(\lambda)}$ or $l_i^{S_1} = y_1^{(\lambda)}$ and $l_j^{S_2} = y_1^{(\lambda)}$, or $l_i^{S_1} = z_1^{(\lambda)}$ and $l_j^{S_2} = z_1^{(\lambda)}$, $i = \overline{1, N_1}$, $j = \overline{1, N_2}$ where N_1, N_2 – amount of all elements of arrays of characteristics in the natural language sentences S_1 and S_2.

For this textual fragment “The USA is the name of the country composed of 50 states joined in a federal republic. It is one of the world’s largest countries” logic and linguistic models are:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)},0, z_1^{(1)},0,0) \,\&\, p_1^{(2)}(x_1^{(2)},0, y_1^{(2)}, q_1^{(2)},0,0,0) \,\&$$
$$p_1^{(3)}(x_1^{(3)}, g_1^{(3)}, y_1^{(3)}, q_1^{(3)},0,0,0).$$

L^{S_1} = is (USA, 0, name, 0, country, 0, 0) &
composed (country, 0, states, 50, 0, 0, 0) &
joined (states, 50, republic, federal, 0, 0, 0).

$$L^{S_2} = p_2^{(1)}(x_2^{(1)},0, y_2^{(1)},0, z_2^{(1)}, r_2^{(1)},0) \,\&\, p_2^{(2)}(x_2^{(2)},0, y_2^{(2)},0, z_2^{(2)}, r_2^{(2)},0).$$

L^{S_2} = is (it, 0, one, 0, countries, world’s, 0) &
is (it, 0, one, 0, countries, largest, 0).

According to this one the previous two rules:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)},0, z_1^{(1)},0,0) \,\&\, p_1^{(2)}(x_1^{(1)},0, y_1^{(2)}, q_1^{(2)},0,0,0) \,\&$$
$$p_1^{(3)}(y_1^{(2)}, g_1^{(3)}, y_1^{(3)}, q_1^{(3)},0,0,0).$$

L^{S_1} = is (USA, 0, name, 0, country, 0, 0) &
composed (USA, 0, states, 50, 0, 0, 0) &
joined (states, 50, republic, federal, 0, 0, 0).

$$L^{S_2} = p_2^{(1)}(x_2^{(1)},0, y_2^{(1)},0, z_2^{(1)}, r_2^{(1)},0) \,\&\, p_2^{(1)}(x_2^{(1)},0, y_1^{(1)},0, z_1^{(1)}, r_2^{(1)},0).$$

L^{S_2} = is (USA, 0, one, 0, countries, world’s, 0) &
is (USA, 0, one, 0, countries, largest, 0).

The outcome of the applying rules is the elements of arrays of characteristics: $l_i^{S_1} = x_1^{(1)}$, $l_{i+1}^{S_1} = z_1^{(1)}$ and $l_j^{S_2} = x_1^{(1)}$, $l_{j+1}^{S_2} = z_1^{(1)}$.

It is possible to use only synonyms for linking phrases and integration the context of different sentences, reiteration of various stylistic interpretations of the one word.

If objects of the relations in two natural language sentences are similar $y_1^{(\lambda)} \equiv y_2^{(\mu)}$ or $\bar{y}_1^{(\lambda)} \equiv \bar{y}_2^{(\mu)}$, where $\bar{y}_1^{(\lambda)} \in \widehat{R}_y$ and $\bar{y}_2^{(\mu)} \in \widehat{R}_y$, that relate to the same range of words with identical root $\widehat{R}_y \in R$ from the set R, or $y_1^{(\lambda)} \in W_y$ and $y_2^{(\mu)} \in W_y$, that relate to the same synonymic range $W_y \subseteq W$ from the set W, so it is necessary to make replacement of $y_1^{(\lambda)} = y_2^{(\mu)}$ into the logic and linguistic models L^{S_1} and L^{S_2} after what:

$$L_{p^{(\lambda)}}^{S_1}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$
$$L_{p^{(\mu)}}^{S_2}(h) = p_2^{(\mu)}(x_2^{(\mu)}, g_2^{(\mu)}, y_1^{(\lambda)}, q_2^{(\mu)}, z_2^{(\mu)}, r_2^{(\mu)}, h_2^{(\mu)}).$$

The elements of arrays of characteristics for both sentences will be: $l_i^{S_1} = y_1^{(\lambda)}$, $i = \overline{1, N_1}$ and $l_j^{S_2} = y_1^{(\lambda)}$, $j = \overline{1, N_2}$, where N_1, N_2 – amount of all elements of arrays of characteristics in the natural language sentences S_1 and S_2.

If subject of the first sentence and object of the second sentence of natural language are identical $x_1^{(\lambda)} \equiv y_2^{(\mu)}$ or $\hat{x}_1^{(\lambda)} \equiv \hat{y}_2^{(\mu)}$, where $\hat{x}_1^{(\lambda)} \in \widehat{R}_x$ and $\hat{y}_2^{(\mu)} \in \widehat{R}_x$, that relate to the same range of words with identical root $\widehat{R}_x \in R$ from the set R, or $x_1^{(\lambda)} \in W_x$ and $y_2^{(\mu)} \in W_x$, that relate to the same synonymic range $W_x \subseteq W$ from the set W, so it is necessary to make replacement of $x_1^{(\lambda)} = y_2^{(\mu)}$ into the logic and linguistic models L^{S_1} and L^{S_2} after what:

$$L_{p^{(\lambda)}}^{S_1}(h) = p_1^{(\lambda)}(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}),$$
$$L_{p^{(\mu)}}^{S_2}(h) = p_2^{(\mu)}(x_2^{(\mu)}, g_2^{(\mu)}, x_1^{(\lambda)}, q_2^{(\mu)}, z_2^{(\mu)}, r_2^{(\mu)}, h_2^{(\mu)}).$$

The elements of arrays of characteristics for both sentences will be: $l_i^{S_1} = x_1^{(\lambda)}$, $i = \overline{1, N_1}$ and $l_j^{S_2} = x_1^{(\lambda)}$, $j = \overline{1, N_2}$, where N_1, N_2 – amount of all elements of arrays of characteristics in the natural language sentences S_1 and S_2.

All this rules consider different versions of the first type of making linkers in textual information - tautological reiteration. This gives an opportunity for computer modelling of the context linkers between the sentences of natural language.

3 Experiment

Let's take textual fragment for searching the context linkers between the sentences of natural language.

According to the rules we have to create logic and linguistic model for each sentence, that consist indivisible content.

Suppose we have such textual information: "Celebrities invite publicity despite knowing that this will leave them open to public attention.

Therefore, it is hypocritical for them to complain when the media shows interest in other aspects of their lives.

Also, celebrities are influential role models to many people and because of this, their private lives should be open to public examination.

Additionally, the public have the right to know about the rich and famous since it is our money that supports them."

Logic and linguistic models for such sentences are:

$$L^{S_1} = p_1^{(1)}(x_1^{(1)},0, y_1^{(1)},0,0,0,0) \rightarrow p_1^{(2)}(0,0, x_1^{(1)}, q_1^{(2)}, z_1^{(2)}, r_1^{(2)},0).$$

$$L^{S_2} = p_2^{(1)}(x_1^{(1)},0,0,0,0,0, h_2^{(1)}) \rightarrow p_2^{(2)}(x_2^{(2)},0, y_2^{(2)},0, z_2^{(2)}, r_2^{(2)},0) \,\&$$

$$p_2^{(3)}(x_2^{(3)},0, y_2^{(3)},0, z_2^{(3)}, r_2^{(3)},0).$$

$$L^{S_3} = x_1^{(1)}(x_3^{(1)}, g_3^{(1)}, y_3^{(1)}, q_3^{(1)},0,0, h_3^{(1)}) \rightarrow p_3^{(2)}(x_3^{(2)}, x_1^{(2)}, y_3^{(2)}, r_1^{(2)},0,0,0) \,\&$$

$$p_3^{(3)}(x_3^{(3)}, x_1^{(1)}, y_3^{(3)}, r_1^{(2)},0,0,0).$$

$$L^{S_4} = p_4^{(1)}(r_1^{(2)},0, y_4^{(1)}, q_4^{(1)}, z_4^{(1)},0, h_4^{(1)}) \,\&\, p_4^{(2)}(r_1^{(2)},0, y_4^{(2)}, q_4^{(2)}, z_4^{(1)},0, h_4^{(1)}) \rightarrow$$

$$p_4^{(3)}(x_4^{(3)}, g_4^{(3)}, x_3^{(2)}, x_1^{(1)}, z_4^{(3)},0, h_4^{(3)}) \,\&\, p_4^{(4)}(x_4^{(4)},0, x_1^{(1)},0,0,0,0).$$

L^{S_1} = invite (celebrities, 0, publicity, 0, 0, 0, 0) →
will_leave (0, 0, celebrities, open, attention, public, 0).

L^{S_2} = complain (celebrities, 0, 0, 0, 0, 0, hypocritical) →
shows (media, 0, interest, 0, aspects, other, 0) &
shows (media, 0, interest, 0, aspects, celebrities_lives, 0).

L^{S_3} = celebrities (role, influential, people, many, 0, 0, models) →
should_be_open (lives, celebrities, examination, public, 0, 0, 0) &
should_be_open (lives, private, examination, public, 0, 0, 0).

L^{S_4} = have (public, 0, right, know, rich, 0, since) &
have (public, 0, right, know, famous, 0, since) →
money (right, know, lives, celebrities, famous, 0, since) &
support (money, 0, celebrities, 0, 0, 0, 0).

The geometric interpretation of these context links between the natural language sentences we can see in Figure 1. There were replaced following components of logic and linguistic models: $x_2^{(1)} = x_2^{(1)}$, $p_3^{(1)} = x_1^{(1)}$ (according to the second rule about similar objects), $y_4^{(4)} = x_1^{(1)}$, $y_4^{(3)} = x_3^{(2)}$ (according to the rule of identical object and subject of the sentence).

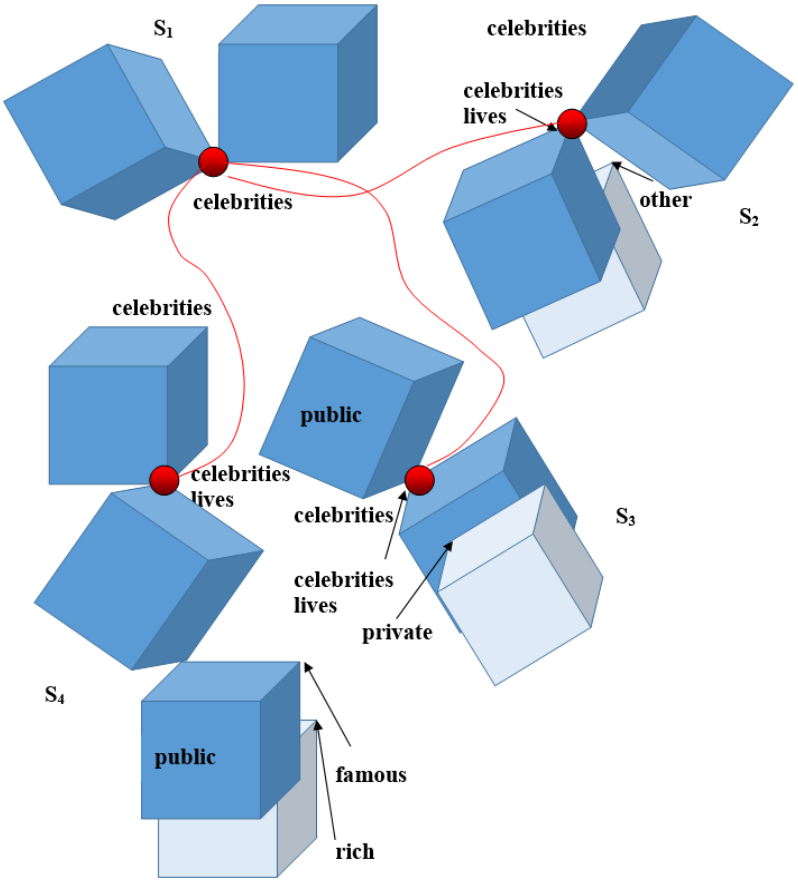

Fig. 1. Geometric interpretation of context links

Every cube interprets simple natural language sentence into the complex sentence. Each corner of the cube – is a component of logic and linguistic model. Grey cubes are sentences with the similar construction and with only one different cube corner. The red lines show context links between four natural language sentences.

According to the rules for searching context links between the natural language sentences author has developed the basic principles for the synthesis of logic and linguistic models of natural language sentences based on the identification of means of meaningful connection (semantic and deictic repetition, use of identical grammatical forms, syntactic or transpositional derivation) in text documents and serves as a basis for constructing logic and linguistic models of electronic text documents. The result is the creation of a knowledge base for information technology of automatic comparative analysis of electronic text documents by content.

It has been made comparative analysis for systems, that are able to show the percentage of matches between electronic textual documents to approve results of finding context links between the natural language sentences. Such systems are as

follows: Advego Plagiatus, Text.ru, StrikePlagiarism, «Антиплагіат» and information technology of automatic comparative analysis with developed knowledge base (СПАТЛЛІМ). It was received the average index for percentage of matches for each system, correspondingly: Advego Plagiatus – 54,55%, Text.ru – 57,98%, StrikePlagiarism – 57,51%, «Антиплагіат» - 51,81% and СПАТЛЛІМ – 69,93% (Figure 2, Figure 3).

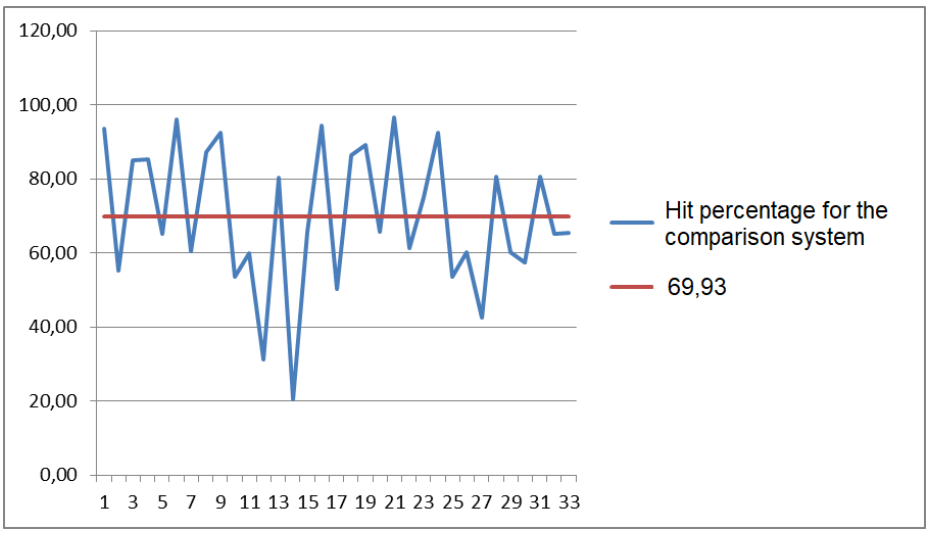

Fig. 2. The average index for percentage of matches for СПАТЛЛІМ

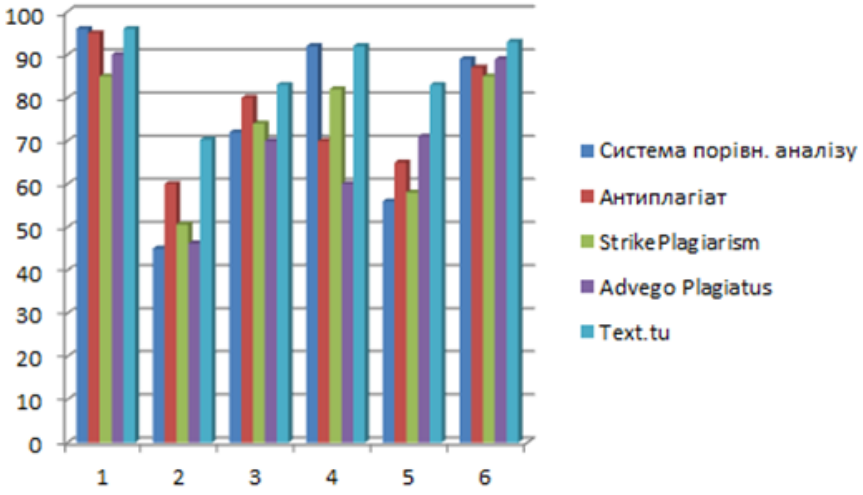

Fig. 3. The statistics of finding the index for percentage of matches

The statistics of finding the index for percentage of matches shows, that information technology which uses the rules for searching context links between natural language sentences, leads to much more better results.

4 Conclusions

The scientific significance of the work lies in using an approach based on the predicate logic method of formation of meaningful models of text documents. It involves the development of a mathematical apparatus for semantical analysis of electronic text documents, which, on the basis of the analysis and synthesis of logical and linguistic models of natural language sentences, enables the structuring of textual information, ranging from the lowest level of logical connections to the text as a whole.

The effectiveness of the proposed rules for searching the context links was shown at geometric interpretation of links between four natural language sentences. According to the proposed rules, the subject of the first sentence connected with the subject of the second, relation of the third and object of the fourth ones. So, solving the problem of knowledge extraction from the textual information can be realized by means of searching semantic reiteration tautological reiteration, thematic reiteration and reiteration of various stylistic interpretations.

References

1. Bargesyan, A., Kupriyanov, M., S., I.I., H.: The technologies of data analys: Data Mining, Visual Mining, Text Mining, OLAP. BHV-Peterburg (2007)
2. Bulgakov, I.: The algorithm of information extraction by ABBYY Compreno (2015), https://habr.com/ru/company/abbyy/blog/269273/
3. Evans, V.: Lexical concepts, cognitive models and meaning-construction. In: Cognitive Linguistics. Vol. 17, pp 73–107. Edinburg university press Publ. (2006)
4. Evans, V., Green, M.: Cognitive Linguistics. Edinburg university press Publ. (2006)
5. Geitgey, A.: Natural Language Processing is Fun (2018), https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e
6. Hlomozda, D., Glybovets, M., Maksymets, O.: Automating the Conversion of Colored Petri Nets with Qualitative Tokens Into Colored Petri Nets with Quantitative Tokens. Cybernetics and Systems Analysis **54**(4), 650–661 (2018). https:// doi.org/10.1007/s10559-018-0066-4
7. Lande, D., Snarskii, A., Yagunova, E., Pronoza, E., Volskaya, S.: Hierarchies of Terms on the Euromaidan Events: Networks and Respondents' Perception. In: 12th International Workshop on Natural Language Processing and Cognitive Science NLPCS 2015 Proceedings. pp. 127–139. Krakow, Poland (2015)
8. Saini, G.: How I used natural language processing to extract context from news headlines (2018), https://towardsdatascience.com/how-i-used-natural-language-processing-to-extract-context-from-news-headlines
9. Shyrokov, V.: The computer lexicography. Naukova dumka (2011)
10. Solveig, B., Bruening, B., Yamada, M.: Affected Experiencers. Natural Language and Linguistic Theory **30**(2), 85–94 (2012). https:// doi.org/10.1007/s11049-012-9177-1

11. Vavilenkova, A.: Basic principles of the synthesis of logical–linguistic models. Cybernetics and systems analysis **51**(5), 826–834 (2015). https:// doi.org/10.1007/s10559-015-9776-z
12. Vavilenkova, A.: Analys and synthesis of logic and linguistic models. TOV "SIK GROUP Ukraine" (2017)