Computer text understanding in a general problem of pattern recognition.

D. V. Mikhailov and G. M. Emelyanov

Yaroslav-the-Wise Novgorod State University

e-mail: *Dmitry.Mikhaylov@novsu.ru*, *Gennady.Emelyanov@novsu.ru*

Subject of research.

Semantic affinity of Natural Language (NL)'s subject-oriented subset's texts.

Research tasks.

1) Formal definition of semantic standard.
2) Development of standards's database's structure for texts's affinity's analysis.
3) Introduction of semantic affinity's measure on the basis of knowledge about situations of semantic equivalence for NL's subset.

The Formal Concept Analysis and situations of natural language usage.

Let's represent the language context formally fixed by NL-usage situation as a triple named the Formal Context (FC):

$$K = (G, M, I), \tag{1}$$

where an objects's set G consists of stems of words being syntactically dependent on other words. An attributes's set M includs subset designated further by M with corresponding bottom index. They contains:

- indications of the syntactically main word's stem. Let's designate further this set as M_1;
- indications for the main word's inflection (M_2);
- «stem–inflection» relations for a main word (M_3);
- combinations of inflections of dependent and main words (M_4). In this case after an inflection a preposition (if any) that provides a relation with a dependent word is shown through a colon;
- indications for dependent word's inflection (M_5).

Definition 1. *A pair (A, B) of sets named as extent (A) and intent (B) forms the Formal Concept (FC) if are true the following mappings:*

$$A' = \{m \in M \,|\, \forall\, g \in A\colon gIm\}\,, B' = \{g \in G \,|\, \forall\, m \in B\colon gIm\},$$

where $A' = B,\ B' = A,\ I \subseteq G \times M$ puts in conformity to objects their attributes.

Definition 2. *A set $\Re(G, M, I)$ of all FCs for K together with the order relation is called the Formal Concept Lattice.*

Definition 3. *A FC of a kind (g'', g') is called the object FC, and a FC of a kind (m', m'') is called the attribute FC, where $g \in G, m \in M$.*

Let $K^E = \left(G^E, M^E, I^E\right)$ be the FC for NL-usage situation S_1 corresponding to the predetermined correct NL-description of a some fact, $K^X = \left(G^X, M^X, I^X\right)$ is a FC for arbitrary NL-usage situation S_2.

We introduce designations for symbol constants: p_{fl} — «флексия:», p_{bs} — «главное-основа:», p_b — «основа:», and symbol $\odot$ for operation of concatenation.

Theorem 1. *Let* $\{m_1, m_2, m_3\} \subset M_1$. *If attributes* m_1, m_2 *and* m_3 *are mutually different then* m_1 *corresponds to the stem's indication for the Splintered Predicative Value (SPV)'s main word;* m_2 *indicates to the SPV's dependent word's stem; and attribute* m_3 *corresponds to the indication for the stem of one-word semantic equivalent of this SPV under necessary fulfillment of the following conditions:*

1) $\exists\, g_1 \in G$: $I(g_1, m_1) = \text{true},\ I(g_1, m_3) = \text{false},\ m_2 = p_{bs} \odot g_1$;

2) $\exists\, \{g_2, g_3\} \subset G$, *thus objects* g_1, g_2 *and* g_3 *are mutually different and*

$$I(g_2, m_3) \wedge I(g_3, m_3) \wedge \\ \wedge \big(I(g_2, m_1) \wedge I(g_3, m_2) \vee \\ \vee\, I(g_2, m_2) \wedge I(g_3, m_1)\big) = \text{true};$$

3) there are no other triples of objects for which the attribute m_3 *occupies the place of either the attribute* m_1 *or* m_2 *in the above relations.*

Remark 1. *After removing SPVs's information, the formal context for the NL-usage situation reflects the classes of relations that are defined exclusively by the roles of objects participating in situation as referred to this proper situation.*

Remark 2. *The words being synonyms can designate the concepts of a different abstract degree. The mentioned degree is higher the larger the number of NL-usage situations relative to which the concept plays a definite role.*

Forming the thesaurus on the basis of NL-usage situations's set.

Let's consider a model of thesaurus in a kind of the formal context:

$$K^H = (G^H, M^H, I^H), \tag{2}$$

where G^H consists of labels of individual NL-usage situations. The set M^H includes the elements of the attributes's sets of FC for all $g^H \in G^H$. In addition, in M^H one can distinguish the following:

- M_6 is the set of indications to an objects of FCs of individual $g^H \in G^H$;
- M_7 is the set of «stem–inflection» relations for a dependent word;
- M_8 is the set of combinations of the stems of the dependent and main words.

The set obtained by uniting the sets $M_6,\ M_7,\ M_8,\ M_4^E,\ M_4^X,\ M_5^E$ and M_5^X will be designated as M^U.

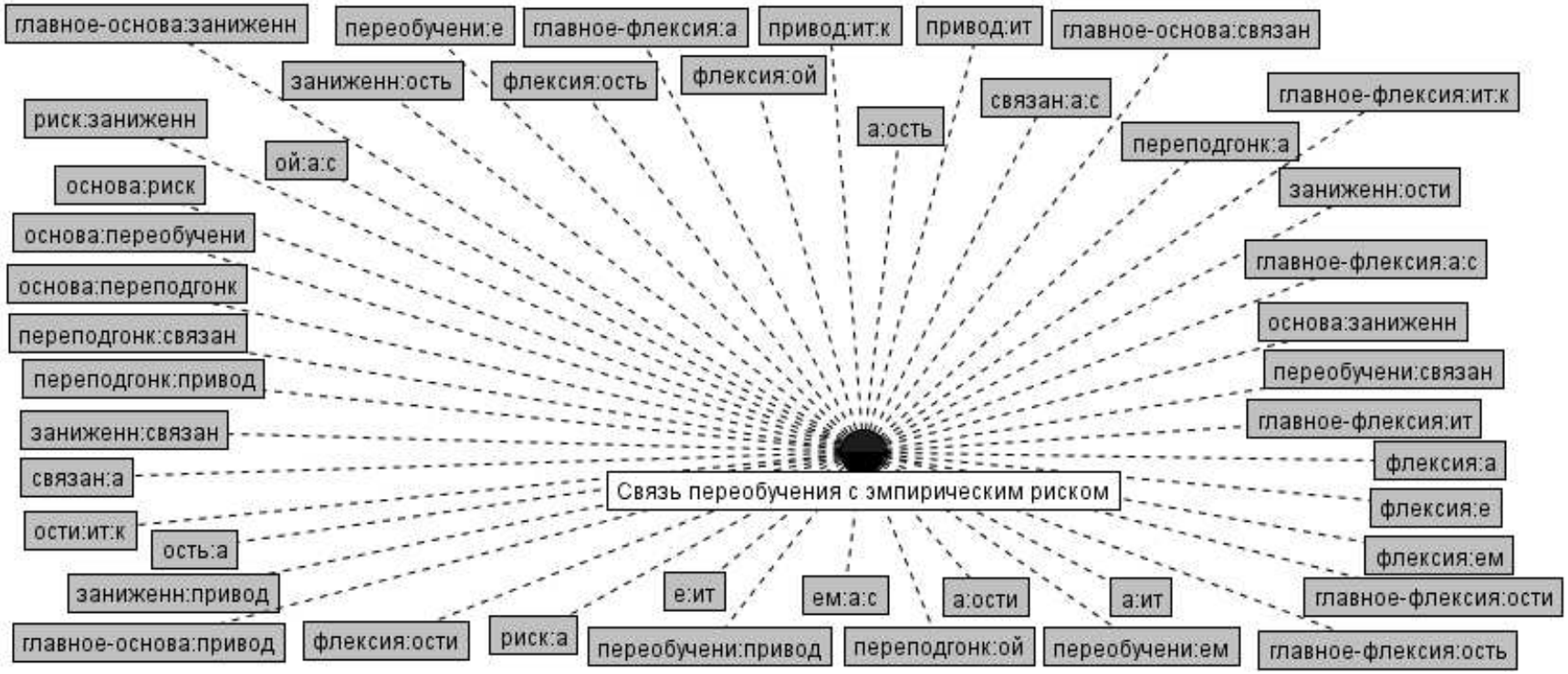

Fig. 1. Object $g^H \in G^H$ for the formal context of individual NL-usage situation.

Definition 4. *Let us take that NL-usage situations S_1 and S_2 are related by the affinity relation if each $g^X \in G^X$ correspond to such $g^E \in G^E$ that one of the following conditions is fulfilled:*
1) $g^X = g^E$ and any attribute $m^E \in M^E$ of object g^E will be related to the object g^X.
2) $g^X = g^E$, Condition 1) is not fulfilled but $g^H \in G^H$ having an attribute $m_1^H \in M_6 \colon m_1^H = p_b \odot g^E$ exists at necessary fulfilment the following conditions:

$$\left(\exists\, m_{fl}^E \in M_5^E \colon m_{fl}^E = p_{fl} \odot f^E\right) \rightarrow \left(\exists\, m_{17}^H \in M_7 \colon m_{17}^H = g^E \odot \text{«:»} \odot f^E\right),$$

$$\text{thus } \left(I^E\left(g^E, m_{fl}^E\right) \wedge I^X\left(g^E, m_{fl}^E\right)\right) \rightarrow I^H\left(g^H, m_{17}^H\right);$$

$$\left(\exists\, m_{bs}^E \in M_1^E \colon m_{bs}^E = p_{bs} \odot b^E\right) \rightarrow \left(\exists\, m_{18}^H \in M_8 \colon m_{18}^H = g^E \odot \text{«:»} \odot b^E\right), \text{thus } I^E\left(g^E, m_{bs}^E\right) \rightarrow I^H\left(g^H, m_{18}^H\right);$$

$$\left(\exists\, m_{bs}^X \in M_1^X \colon m_{bs}^X = p_{bs} \odot b^X\right) \rightarrow \left(\exists\, m_{28}^H \in M_8 \colon m_{28}^H = g^E \odot \text{«:»} \odot b^X\right), \text{thus } I^X\left(g^E, m_{bs}^X\right) \rightarrow I^H\left(g^H, m_{28}^H\right).$$

In addition, for $\forall\, m^H \in (M^H \backslash M^U)$ is true the following implication:

$$I^H\left(g^H, m^H\right) \rightarrow \left(I^E\left(g^E, m^H\right) \wedge I^X\left(g^E, m^H\right)\right). \tag{3}$$

3) $g^X \neq g^E$, but there is $g^H \in G^H$ having attributes $m_1^H \in M_6 \colon m_1^H = p_b \odot g^E$ and $m_2^H \in M_6 \colon m_2^H = p_b \odot g^X$, thus for any $m^H \in (M^H \backslash M^U)$ is true the following:

$$I^H\left(g^H, m^H\right) \rightarrow \left(I^E\left(g^E, m^H\right) \wedge I^X\left(g^X, m^H\right)\right). \tag{4}$$

4) $g^X \neq g^E$, $\exists\,(g_1^H \in G^H, m_1^H \in M_6) : I^H\left(g_1^H, m_1^H\right) = \text{true}$, $m_1^H = p_b \odot g^E$ and for $\forall\, m^E \in (M_4^E \cup M_5^E)$ $\left(I^H\left(g_1^H, m_1^H\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^H\left(g_1^H, m^E\right)$ is true. In this case there are attributes $m_2^H \in M_6$ and $m^X \in (M_1^X \cup M_2^X \cup M_3^X)$ for which $\left(I^H\left(g_1^H, m_2^H\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^H\left(g_1^H, m^X\right)$ is true, where $m_2^H = p_b \odot g^{X_1}$, $g^{X_1} \neq g^X$ and the pair (g^{X_1}, g^E) meets Condition 3) of the present Definition in generation of the formal context for the object g_1^H. At the same time there is an object $g_2^H \in G^H$ concerning which the pair (g^X, g^{X_1}) also meets Condition 3) of the present Definition. We designate the generated formal context for object g_2^H as K^{X_1}. By analogy with K^E and K^X $K^{X_1} = (G^{X_1}, M^{X_1}, I^{X_1})$.

Affinity's measure for NL-usage situations.

The affinity's measure for NL-usage situations S_1 and S_2 relatively to FCs $K^E = (G^E, M^E, I^E)$ and $K^X = (G^X, M^X, I^X)$ from which an information of SPVs is removed, is calculated as:

$$spc(S_1, S_2) = \frac{\sum_{k=1}^{n} spc_k}{n}, \tag{5}$$

where $n = |G^X|$ and spc_k is the objects's affinity's measure in the pair (g_k^X, g^E). The value of spc_k:
- equals 1.0 if for the pair (g_k^X, g^E) *Condition 1)* in *Definition 4* is fulfilled;
- is calculated by the formula:

$$-\log_2\left(1 - \frac{D_c}{path_C}\right) \times \frac{|B^C|}{|B_1 \backslash B^C| + |B_2 \backslash B^C| + |B^C|}, \tag{6}$$

if for the pair (g_k^X, g^E) *Condition 2), 3) or 4)* in *Definition 4* is fulfilled.

If $\exists\, g^X \in G^X$ for which there is no feasible conditions in *Definition 4* then $spc(S_1, S_2) = 0$.

In case of fulfilment of any of *Conditions 2)–4)* in *Definition 4*, $D_c = 2$ (the proof is evident).

If *Condition 2)* or *3)* is fulfilled, $path_C = 4$ and the set B^C will include attributes $m^H \in (M^H \backslash M^U)$ for each of them either meet relation (3) (if *Condition 2)* is fulfilled) or meet relation (4) (if *Condition 3)* is fulfilled). In this case the sets B_1 and B_2 are determined as follows:

$$B_1 = \left\{ m^E : m^E \in \left(M_1^E \cup M_2^E \cup M_3^E \right), I^E\left(g^E, m^E\right) = \text{true} \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^X \cup M_2^X \cup M_3^X \right), I^X\left(g_k^X, m^X\right) = \text{true} \right\}.$$

The feasibility of *Condition 4)* is commonly proved by several iterations. In each subsequent iteration, the number of attributes being uncommon for g_k^X and g^{X_1} is always fewer than that in the previous iteration. The initial value $path_C = 4$ increases by 1 in each iteration. If the *Condition 4)* is true then

$$B_1 = \left\{ m^{X_1} : m^{X_1} \in \left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1} \right), I^{X_1}\left(g^{X_1}, m^{X_1}\right) = \text{true} \right\},$$
$$B_2 = \left\{ m^X : m^X \in \left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1} \right), I^{X_1}\left(g_k^X, m^X\right) = \text{true} \right\},$$

where $\left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right) \subset M^{X_1}$. The set B^C here is the intersection of B_1 and B_2.

Table 1. Initial data for thesaurus building (in Russian).

№п/п	1				2	3		4	
stem	inflection + preposition								
заниженн	ость	ость	ости	ости	—	ость	ости	ость	ость
оценк	—	—	—	—	—	и	и	и	и
эмпирическ	ого	—	ого	—	—	—	—	—	—
риск	а	—	а	—	—	—	—	—	—
средн	—	ей	—	ей	—	—	—	—	—
ошибк	—	и:на	—	и:на	—	—	—	и	и
распознавани	—	—	—	—	—	—	—	я	я
обучающ	—	ей	—	ей	—	—	—	—	—
выборк	—	е	—	е	—	—	—	—	—
переусложнени	ем	ем	е	е	—	—	—	—	—
модел	и	и	и	и	—	—	—	—	—
уменьшени	—	—	—	—	е	—	—	—	—
обобщающ	—	—	—	—	ей	ей	ей	—	—
способност	—	—	—	—	и	и	и	—	—
выбор	—	—	—	—	—	—	—	ом	а
решающ	—	—	—	—	его	—	—	его	его
дерев	—	—	—	—	а	—	—	—	—
правил	—	—	—	—	—	—	—	а	а
алгоритм	—	—	—	—	—	а	а	—	—
переподгонк	—	—	—	—	ой	ой	а	—	—
переобучени	—	—	—	—	—	ем	е	—	—
связан	а:с	а:с	—	—	о:с	а:с	—	а:с	—
вызван	а	а	—	—	—	а	—	—	—
обусловлен	а	а	—	—	о	—	—	—	—
привод	—	—	ит:к	ит:к	—	—	ит:к	—	—
завис	—	—	—	—	—	—	—	—	ит:от

Example of initial data for formal contexts's building for the compared NL-usage situations.

Table 1. Russian descriptions of the relation between overfitting and empirical risk.

NL-description	standard				analyzed			
variant	1	2	3	4	1	2	3	4
stem	inflection + preposition							
заниженн	ости	ости	ость	ость	ость	ость	ости	ость
эмпирическ	ого	ого	ого	ого	—	—	—	ому
риск	а	а	а	а	—	—	—	у
средн	—	—	—	—	ей	ей	ей	ей
ошибк	—	—	—	—	и:на	и:на	и:на	и:на
обучающ	—	—	—	—	ей	ей	ей	ей
выборк	—	—	—	—	е	е	е	е
переобучени	е	—	—	ем	ем	—	е	—
переподгонк	—	а	ой	—	—	ой	—	—
связан	—	—	а:с	а:с	а:с	а:с	—	—
привод	ит:к	ит:к	—	—	—	—	ит:к	ит:к

Comment. *Variant No 4 of analyzed Russian description of considered fact is incorrect: «Заниженность средней ошибки на обучающей выборке приводит к эмпирическому риску».*

Result : value of affinity to the standard for analyzed variants of given subject area's fact.

Table 2. Comparison of the variants of NL-description of the relation between overfitting and empirical risk.

Variant	$spc(S_1, S_2)$	$\left\|B^C\right\|$	$\left\|B_1 \backslash B^C\right\|$	$\left\|B_2 \backslash B^C\right\|$
1	0.9167	7.7500	0.7500	0.0000
2	0.7917	7.0000	2.0000	0.5000
3	0.8750	7.7500	0.7500	0.7500
4	0.0000	—	—	—

Conclusions.

- The main *result* of the present work is *the method for analysis of mutual affinity of natural language's usage situations in their independent generation*. An application of formal concept lattices to present the compared NL-phrases and the thesaurus information allows for easy replenishment of thesaurus and effective usage of information available in analysis of text affinity.
- The proposed *thesaurus model* can be used as the basis for building standards's database for a specified subject area. Owing to hierarchical presentation of information in the formal concept lattice the size of standards's database and search time in it can be reduced.
- A thesaurus's model in the form of formal concept lattice ensures the informational compression either at the expense (first of all) of predicate words which designate situations are similar to some extent by membership of participants and type of their actions, or at the expense of abstract lexicon. Informational compression's degree depends on relevance to the specified subject area of each of the separate facts's descriptions presented in lattice.
- Separate applied research is required for quantitative estimations of completeness of coverage of the language description of subject knowledge in a thesaurus lattice.