# СМЫСЛОСОХРАНЯЮЩЕЕ СЖАТИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ ЗНАНИЙ О СИНОНИМИИ[1]

**Д. В. Михайлов, Г. М. Емельянов[2]**

**[2] Новгородский государственный университет имени Ярослава Мудрого**
**173003, Россия, Великий Новгород, ул. Большая С.-Петербургская, 41, тел.:(8162)627940**
**e-mail: Dmitry.Mikhaylov@novsu.ru, Gennady.Emelyanov@novsu.ru**

Статья посвящена проблеме передачи текстовой информации с минимальными потерями полезной составляющей. Предложено её решение в рамках теории анализа формальных понятий на основе ситуации языкового употребления как единицы формализованного описания семантики.

## Введение

В настоящее время одной из основных задач интеллектуальных систем является накопление предметных знаний и их обмен между людьми. Немаловажную роль при этом играют знания, представляемые текстами на Естественном Языке (ЕЯ).

К примеру, для интерпретации результатов теста открытой формы в системе контроля знаний необходимо учитывать различные формы описания одного и того же факта действительности разными экспертами на одном и том же ЕЯ. При этом актуальна задача поиска наиболее рационального плана передачи смысла, сам же смысл в итоге должен быть отражён в максимально компактном объёме текстовых данных. Именно эти данные участвуют в оценке близости ответа испытуемого правильному ответу. Данная работа посвящена решению указанной задачи на основе предложенной авторами концепции смыслового эталона Ситуации Языкового Употребления (СЯУ).

## Ситуация языкового употребления

Пусть $Ts$ есть множество Семантически Эквивалентных (СЭ) ЕЯ-фраз, задающих различные формы описания некоторого факта действительности и определяющих СЯУ. Представим СЯУ посредством тройки

$$K = (G, M, I), \qquad (1)$$

именуемой в теории анализа формальных понятий [1] Формальным Контекстом (ФК). Множество его объектов $G$ составляют основы слов, синтаксически зависимых по отношению к другим словам из СЭ-фраз в составе $Ts$, признаковое множество $M$ включает признаки-указания на основы и флексии слов, синтаксически главных по отношению к словам с основами из множества $G$. В множество $M$ входят также связи «основа–флексия» для главного слова и комбинации флексий зависимых и главных слов.

Ставится задача сформировать $I \subseteq G \times M$ анализом буквенной структуры фраз из $Ts$. Основу формирования $I$ при этом должны составить фразы, отвечающие требованию компактности выражения смысла.

## Смысловой эталон и его применение

В задачах классификации под гипотезой компактности понимают предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в различных [2]. Если представить смысл множества фраз $\{Ts_i : Ts_i \in Ts\}$ как набор функций, которые связывают обозначаемые словами понятия, то каждая такая функция:

- задаётся на множестве буквенных цепочек, составляющих основы слов фраз $Ts_i \in Ts$;
- имеет множество значений, однозначно определяемое некоторым $I' \subset I$.

Компактное представление смысла здесь означает минимизацию символьной длины

[1] Работа поддержана РФФИ, проект № 13-01-00055

$Ts_i$ при максимизации числа слов $w_j \in Ts_i$, наиболее употребимых в различных фразах из $Ts$ (с учётом возможных синонимов).

Обозначим далее множество индексов для неизменных частей (основ) слов фраз из $Ts$ как $J$. Последовательность таких индексов для некоторой $Ts_i \in Ts$ назовём моделью её линейной структуры (МЛС), $Ls(Ts_i)$.

Пусть $LS$ – множество моделей линейных структур фраз из $Ts$ на $J$.

**Лемма 1.** Пара $\{j_1, j_2\} \subset J$ соответствует синонимам, если $\exists \{Ls(Ts_1), Ls(Ts_2)\} \subseteq LS$:

$$\begin{cases} Ls(Ts_1) = J_{bef} \bullet \{j_1\} \bullet J_{aft} \\ Ls(Ts_2) = J_{bef} \bullet \{j_2\} \bullet J_{aft} \end{cases},$$

где $J_{bef} \subset J$, $J_{aft} \subset J$, а "$\bullet$" есть операция типа конкатенации над $J$.

Пусть $PJ$ – множество пар, отвечающих Лемме 1. Заменим индексы, вошедшие в пары из $PJ$, на некоторые $j \in (\mathrm{N} \setminus J)$ во всех моделях из $LS$. Обозначим далее преобразованное множество $LS$ как $LS'$.

**Утверждение 1.** Пусть $\{J_1, J_2\}$ – пара последовательностей индексов в $Ls(Ts_i)$, где $J_1 = \{j_1^1, \ldots, j_2^1\}$, $J_2 = \{j_1^2, \ldots, j_2^2\}$, а каждой из пар $(j_1^1, j_2^1)$ и $(j_1^2, j_2^2)$ отвечает синтаксическая связь. Тогда смысловой эталон СЯУ определяют те $Ts_i \in Ts$, в МЛС которых

$$\begin{cases} J_1 \subset J_2 \\ J_2 \subset J_1 \\ |J_1 \cap J_2| = 1 \\ J_1 \cap J_2 = \varnothing \end{cases}, \qquad (2)$$

а сумма длин таких последовательностей для всех связей на $Ts_i$ минимальна.

**Утверждение 2.** Пусть $fr(w_j)$ – частота появления слова $w_j$ (независимо от его формы) во всех $Ts_i \in Ts$. Основу эталона составляют фразы с максимумом слов, вошедших в особый кластер $Clust$:

- слово с максимальным значением этой частоты войдёт в $Clust$;
- для $\forall \{w_j, w_k\} \subset Clust$ и $\forall w_l \notin Clust$ верно то, что

$$\begin{cases} |fr(w_j) - fr(w_k)| < |fr(w_j) - fr(w_l)| \\ |fr(w_j) - fr(w_k)| < |fr(w_k) - fr(w_l)| \end{cases}.$$

**Замечание.** При формировании множества $Clust$ учитываются возможные синонимы анализируемых слов (согласно Лемме 1), поэтому для любого $w_j$ значение $fr(w_j)$ оценивают относительно множества $LS'$.

Пусть $J_{Cl} \subset J$ – множество индексов слов, вошедших в $Clust$. Рассмотрим множество $LC = \bigcup_i LS_i : LS_i \subset LS, \exists Ts_i, Ts_j \in Ts$:

$$Ls(Ts_i) \in LS_i,$$
$$|Ls(Ts_i) \cap J_{Cl}| \to \max,$$
$$\left((Ls(Ts_j) \in LS_i) \wedge (Ts_j \neq Ts_i)\right) \to$$
$$\to (Ls(Ts_i) \cap J_{Cl}) \subset Ls(Ts_j).$$

Как следует из Утверждения 2, смысловой эталон определяют те фразы, МЛС которых принадлежат $LC$.

Для построения признакового множества ФК вида (1) эталона СЯУ требуется найти индексные пары, отвечающие условию (2), и каждой паре нужно задать направление соответствующей синтаксической связи.

**Алгоритм 1.** Формирование связей.

**Вход:** $LS$;

**Выход:** $R_J = \{((j,k), Dir) : Dir \in \{\leftarrow, \rightarrow\}\}$;

**Начало**

1: $R_J := \varnothing$;
2: сформировать $LC$ на основе $LS$;
3: **для всех** $Ls(Ts_i) \in LC$
4: $P_i := \{(j,k) : j,k \in Ls(Ts_i), j \neq k\}$;
5: $P := \bigcup_i P_i$ с учётом $(j,k) \Leftrightarrow (k.j)$;
6: $P' := \{(j,k) \in P : frq((j,k), LS) > 1\}$;
7: **для всех** $(j,k) \in P'$
8: **если** найдено $Dir(j,k)$ **то**
9: $R_J := R_J \cup \{(j,k), Dir\}$;

**Конец** *{Алгоритм 1}*.

Здесь $frq((j,k), LS)$ есть частота появления пары $(j,k)$ в моделях из множества $LS$ с учётом того, что $(j,k) \Leftrightarrow (k.j)$.

Для каждой пары $(j,k)$ из выделенных на Шаге 6 Алгоритма 1 поиск $Dir(j,k)$ идёт в три этапа. На первом проверяется, является ли связь, соответствующая паре, ложной.

[1] Работа поддержана РФФИ, проект № 13-01-00055

**Определение 1.** Пусть $\{j,k,l\} \subset J$, а $St(j)$, $St(k)$ и $St(l)$ есть основы слов, отвечающие индексам $j$, $k$ и $l$. Связь, ассоциируемая с парой $(j,k)$, идентифицируется как ложная относительно рассматриваемой СЯУ при одновременном выполнении условий:

1. $\exists Ts_i \in Ts:\ j,k,l \in Ls(Ts_i)$.
2. В рассматриваемой предметной области существует СЯУ, где связь между $St(j)$ и $St(k)$ идентифицирована как ложная, но существует связь либо между $St(j)$ и $St(l)$, либо между $St(k)$ и $St(l)$.

**Замечание.** Начальные знания системы об истинных и ложных связях формируются в режиме интервью с экспертом. При этом совокупным знаниям по отдельной СЯУ соответствует булев вектор

$$(d_1,\ldots,d_k,\bar{d}_{k+1},\ldots,\bar{d}_n),$$

где компоненты $d_1,\ldots,d_k$ отождествляются с истинными, а $\bar{d}_{k+1},\ldots,\bar{d}_n$ – с ложными связями.

Пару $(j,k)$, доказать ассоциацию с ложной связью для которой не удалось, проверяют на возможность отождествления с ранее выделенными связями.

Пусть $w(j) \in Ts_i$: $w(j) = St(j) \bullet Fl(j)$, где символьная цепочка $Fl(j)$ представляет флективную часть слова $w(j)$, а символом «$\bullet$» обозначается операция конкатенации. Аналогично пусть $w(k) \in Ts_i$ и при этом $w(k) = St(k) \bullet Fl(k)$. Обозначим множество ранее выделенных связей как $Lnk$. Каждый элемент в $Lnk$ представляется четвёркой

$$(Id, St_1, St_2, FCm),$$

где $Id$ – идентификационный номер СЯУ; $St_1$ – основа главного, $St_2$ – зависимого слова; $FCm$ – список пар вида «флексия главного слова – флексия зависимого».

Считается, что паре $(j,k)$ соответствует связь $((j,k),\rightarrow)$ в рамках заданной СЯУ, если для некоторой СЯУ с номером $Id$

$$\exists (Id, St_1, St_2, FCm) \in Lnk:\ St(j) = St_1,$$
$$St(k) = St_2, \text{ а } (Fl(j), Fl(k)) \in FCm.$$

В случае, когда $St(j) = St_2$, $St(k) = St_1$, а $FCm$ содержит пару $(Fl(k), Fl(j))$, паре $(j,k)$ будет отвечать связь $((j,k),\leftarrow)$.

Как и на этапе формирования начальных знаний, пару $(j,k)$, для которой не нашлось ассоциации ни с одной из уже выделенных связей (ложных или истинных), проверяют на наличие связи, опрашивая эксперта.

На основе найденного $R_J$ далее идёт отбор фраз $Ts_i \in Ts$ для построения множества признаков ФК (1) эталона СЯУ.

Первым шагом из состава $\forall LS_i \subset LC$ исключаются те МЛС, которые включают индексы, не вошедшие ни в одну из связей в составе $R_J$. Введём обозначение $LC^*$ для преобразованного таким образом $LC$, аналогично $LS_i^*$ – для $\forall LS_i \subset LC$.

По каждому $LS_i^* \subset LC^*$ отбирается $Ts_i$:

$$Ls(Ts_i) \in LS_i^*,\ |Ts_i| \rightarrow \min. \qquad (3)$$

Совокупность фраз $Ts_i \in Ts$, отвечающих условию (3), обозначим как $Ts^*$.

Заключительный шаг формирования ФК вида (1) эталона СЯУ состоит в построении признакового множества $M$ и объектно-признаковых связей в рамках отношения $I \subseteq G \times M$ на основе найденных $R_J$ и $Ts^*$.

В целях более точного выделения объектов и признаков эталона введём процедуру согласования знаний относительно разных СЯУ заданной предметной области.

Пусть модель (1) есть единица тезауруса, представляемого тройкой

$$Kth = (Gth, Mth, Ith), \qquad (4)$$

где $Gth$ состоит из символьных пометок отдельных СЯУ, $Mth$ включает признаки ФК вида (1) каждой $gth \in Gth$. Кроме того, в $Mth$ входят указания на объекты ФК (1) для $gth \in Gth$, сочетания основы и флексии зависимого, основ зависимого и главного слова. Модель (4) позволяет определить процедуру согласования единиц знаний с помощью следующего правила.

**Правило 1.** Пусть $St_j$ есть основа, $Fl_j$ – флексия слова $w$, найденные относительно СЯУ $S_j$. Предположим, что $w = St_1 \bullet Fl_1$ для СЯУ $S_1$, $w = St_2 \bullet Fl_2$ для СЯУ $S_2$, причём $St_1 = St_2 \bullet suf$, где $suf$ содержит минимум один символ. Тогда относительно

[1] Работа поддержана РФФИ, проект № 13-01-00055

$S_1$ основа $St_1$ может быть заменена на $St_2$, а флексия $Fl_1$ – на $Fl_3 = suf \bullet Fl_2$ только в том случае, если встречаемость флексий $Fl_3$ и $Fl_2$ в отношениях из $Ith \subseteq Gth \times Mth$ не уменьшится при выполнении этих замен. Так, для СЯУ из табл. 1 согласование их эталонов как единиц тезауруса, задаваемого моделью (2), дополнительно сокращает его размер в среднем на 1,5%. Для сравнения в табл. 2 приводятся значения числа СЭ-фраз, задающих СЯУ ($N_1$), фраз, определяющих эталон ($N_2$), исходного числа объектов ($N_3$) и признаков СЯУ ($N_4$), числа объектов ($N_5$) и признаков эталона ($N_6$).

**Таблица 2. Эталоны для СЯУ из табл. 1**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $N_1$ | 56 | 28 | 29 | 30 | 6 | 10 |
| $N_2$ | 8 | 9 | 7 | 9 | 1 | 2 |
| $N_3$ | 18 | 17 | 15 | 13 | 12 | 14 |
| $N_4$ | 177 | 186 | 173 | 162 | 94 | 81 |
| $N_5$ | 9 | 12 | 12 | 11 | 8 | 12 |
| $N_6$ | 82 | 90 | 80 | 69 | 35 | 53 |

**Таблица 1. Ситуации языкового употребления**

| $i$ | Фраза максимальной длины из определяющих СЯУ |
|---|---|
| 1 | *Нежелательное переобучение является причиной заниженности средней величины ошибки алгоритма на обучающей выборке.* |
| 2 | *Тренировочная выборка, на ней проявляется эффект заниженных значений средней ошибки, причиной же является переусложненная модель.* |
| 3 | *Контрольная выборка, принятие деревом решения на ней будет с большей вероятностью ошибки именно по причине переподгонки.* |
| 4 | *Оценка частоты ошибок на выборке, взятой в качестве контрольной, может для алгоритма оказаться заниженной по причине переподгонки.* |
| 5 | *Заниженность оценки ошибки распознавания зависит от выбора правила принятия решений.* |
| 6 | *Число закономерностей алгоритмической композиции влияет на частоту ошибок логического классификационного алгоритма на контрольной выборке.* |

**Таблица 3. Оценка объёма памяти для хранения ЕЯ-фразы**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n$ | 12 | 15 | 16 | 17 | 10 | 14 |
| $vol(n)$ | $4.790 \cdot 10^8$ | $1.308 \cdot 10^{12}$ | $2.092 \cdot 10^{13}$ | $3.557 \cdot 10^{14}$ | $3.629 \cdot 10^6$ | $8.718 \cdot 10^{10}$ |
| $vol_1(n)$ | 648 | 795 | 416 | 442 | 20 | 42 |
| $vol_2(n)$ | 168 | 225 | 80 | 187 | 20 | 42 |

Рассмотренная концепция СЯУ позволяет оценить объём памяти для хранения текстов. Традиционно такой оценкой для фразы из $n$ слов берут $vol(n) = n!$. Введение же эталона СЯУ позволяет дать оценку сверху как $vol_1(n) = l_1 \cdot n$ и снизу как $vol_2(n) = l_2 \cdot n$, где $l_1$ – число СЭ-фраз из задающих СЯУ, из которых $l_2$ определяют эталон.

Соотношение указанных оценок для СЯУ из табл. 1 представлено в табл. 3.

## Заключение

Предложенный метод выделения эталона СЯУ реализован в рамках демо-версии системы контроля знаний, представленной (с исходными текстами на Visual Prolog 5.2) в подразделе «Участник:Dmitry.Mikhaylov» раздела «Страницы участников» ресурса [2]. Согласование знаний по Правилу 1 аналогично самоорганизации смысла слов в мультиагентном подходе [3], что позволяет на основе модели (1) находить системы зависимостей совместной встречаемости осмысленных фрагментов слов и сокращать перебор при построении модели контекста.

[1] Работа поддержана РФФИ, проект № 13-01-00055

## Список литературы

1. Ganter B. and Wille R. Formal Concept Analysis - Mathematical Foundations // Berlin: Springer-Verlag, 1999.
2. http://www.machinelearning.ru (дата обращения: 27.04.2013).
3. Минаков И.А. Интеграция профессиональных знаний, представленных в виде текстов на естественном языке // Вестник СамГТУ, серия «Технические науки» , 2007. № 1(19). С.28–35.