Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности.

Д. В. Михайлов, Г. М. Емельянов

Новгородский Государственный Университет имени Ярослава Мудрого

Цель.

Разработка и исследование методов формирования прецедентов для классов Семантической Эквивалентности (СЭ) высказываний Естественного Языка (ЕЯ).

Задачи исследования.

1) Разработка и исследование методик формализации смысла слова как характеристики его Лексического Значения (ЛЗ) для заданного класса СЭ.
2) Построение формального аппарата математического моделирования процесса обобщения независимых формализованных описаний толкований ЛЗ слова как основы выделения его смысла.
3) Разработка математической модели процесса выявления и обобщения Смыслового Отношения (СО) в рамках Расщепленного Значения при формировании прецедентов СЭ для ситуаций использования Лексических Функций-параметров.
4) Разработка и совершенствование методов анализа корректности толкований ЛЗ как исходных данных для формирования прецедентов.
5) Апробация полученных методик для известных в лингвистике классификаций СЭ.

Формирование классов СЭ для ситуаций синонимии в рамках стандартных Лексических Функций.

Дано :

Π^R — множество правил синонимических преобразований ЕЯ-высказываний;

L^{Π} — множество пар ЕЯ-высказываний, между которыми возможно установление синонимии (относительно Π^R);

$r(\pi)$ — условие применимости $\pi \in \Pi^R$. Для $T = \{T_1, T_2\} : T \in L^{\Pi}$ $r(\pi)$ есть совокупность требований к $\forall w_i \in W$, $W = W_1 \cup W_2$, $W_1 \subset T_1$, $W_2 \subset T_2$, W_1 и W_2 — совокупности слов, заменяемых посредством π;

Lm^W — множество формализованных толкований ЛЗ слов $w_i \in W$ в виде теорий.

Требуется :

- Применительно к $\forall r(\pi)$ выделить множество значимых признаков для $\forall w_i \in W$ анализом $Lm(w_i)$, $Lm(w_i) \in Lm^W$;
- На основе выделенных признаков либо отнести предъявляемую произвольную пару T к одному из известных классов $\pi \in \Pi^R$ СЭ, либо образовать с помощью T новый класс.

Ситуация СЭ на основе расщепленного значения.

Определение 1. *Расщепленное Значение (РЗ) :*

- *Описывается комплексом* $W_j, j \in \{1,2\}$ *лексических единиц :* $\forall w_i \in W_j$ *либо является значением некоторой Лексической Функции (ЛФ)* F *для ключевого слова* $C0$*, определяющего ситуацию СЭ, либо есть само* $C0$*;*
- $\exists w_i \in W_j : w_i = F_i(C0)$ *и* F_i *относится к классу Лексических Функций-параметров;*
- *Может быть выражено одним словом, представляющим собой значение некоторой ЛФ-замены для данного* $C0$*, либо само* $C0$*.*

Замечание. *Формирование* $r(\pi)$ *для ситуации СЭ на основе РЗ предполагает наряду с формализацией требований к смыслу слов из* W*, выявление и обобщения Смыслового Отношения между* $\forall w_i \in W$ *и* $\forall w_m \in W, w_m \neq w_i$ *:*

- $w_i = F_i(C0)$*, где* F_i *— некоторая ЛФ-параметр для заданного* $C0$*;*
- $w_m = F_m(C0)$*, где* F_m *— некоторая ЛФ-замена для заданного* $C0$*, либо* $w_m = C0$*.*

Пример. *РЗ «осуществлять эксперимент», где значением ЛФ* $Oper_1$ *задается СО типа «операция с. . . » между 1-м участником ситуации СЭ (кто осуществляет эксперимент) и ее названием («эксперимент»). Данное РЗ эквивалентно ЛЗ «экспериментировать».*

Смысл как набор формальных атрибутов Лексического Значения.

Пусть для $\forall w_i \in W_j$, $W_j \subset T_j$, $T_j \in T$, $T \in L^{\Pi}$ имеется описание теории ЛЗ :

$$Lm(w_i) = (w_i, L^M), \tag{1}$$

совокупностью бинарноых отношений R_2 между понятиями C_1 и C_2 :

$$M_p = (R_2, C_1, C_2), \tag{2}$$

а также рекурсивно определяемых отношений произвольной арности :

$$M'_p = (R_n, C, L^M) \text{ и} \tag{3}$$

$$M''_p = (R_c, L^M) \tag{4}$$

Здесь L^M — список структур (2), (3) и (4), $R_c \in \{\vee, \&, \neg\}$. Посредством L^M в (3) задается связь понятия C с другими словами и понятиями.

Определение 2. *Если* $\exists Lm(w_i) = (w_i, L^M)$ *(1), то смысл* $\forall w_i \in T$ *определяется набором Характеристических Функций (ХФ)* $\{ChF_{hi}\}$ *:*

- $\exists M_p = (R_2, C_1, C_2) =: ChF_{Val} : M_p \in L^M$ *(2),* $ChF_{hi}(w_i) = C_2 : C_2$ *— обозначение известного системе понятия (Семантического Класса (СК)).* L^M *может быть третьим аргументом (3);*
- $\exists M_p = (ChF_{hi}, C'_1, C'_2) =: ChF_{Name}$*, либо* $\exists M'_p = (ChF_{hi}, C, L^M) =: ChF_{Name} : ChF_{hi}$ *— имя известного СК или СО;*
- ChF_{Name} *— первое из удовлетворяющих вышеуказанному условию при обратном просмотре* L^M *от* ChF_{Val}*. Обозначим* $L^{M\prime} \in L^M$ *: либо* $L^{M\prime} = \{(ChF_{hi}, C'_1, C'_2), \ldots, (R_2, C_1, ChF_{hi}(w_i))\}$*, либо* $L^{M\prime} = \{(ChF_{hi}, C, L^M)\}$*,* $M_p = (R_2, C_1, ChF_{hi}) \in L^M$*;*
- *каждое последующее утверждение в* $L^{M\prime}$ *должно иметь как минимум один общий аргумент, являющийся обозначением некоторой переменной, с предыдущим утверждением.*

Пример формирования набора ХФ для заданного ЛЗ.

Рис. 1. Анализируемый вариант теории ЛЗ

Рис. 2. Характеристические Функции и формальные признаки их значений

Здесь $Var\_SomeBody$ обозначает переменную для слова, интерпретируемого посредством (1). Она же является вторым аргументом для ChF_{Name}.

Модель системы независимых теорий Лексического Значения.

$$K = (G, M, V, I) \tag{5}$$

Здесь :

G — множество объектов. $\forall g \in G : g = Lm_j(w_i)$ есть j-й вариант толкования ЛЗ w_i в форме (1);

M — множество признаков. $M = M_1 \cup M_2$, где $\forall m \in M_1 : m = ChF_{hi}(w_i)$, а $\forall m \in M_2 : \exists ChF_{hi}(w'_i) : Lm(w'_i) = (w'_i, L^{M\prime}) : \exists (R''_2, C''_1, C''_2) \in L^{M\prime}$, где R''_2 — имя известного СК или СО, причем $m = R''_2$;

V — множество значений признаков. $V = V_1 \cup V_2$, где $\forall v \in V_1$ есть имя Характеристической Функции ChF_{hi}, причем задано ее значение $ChF_{hi}(w_i)$ для w_i, а $\forall v \in V_2$ есть значение $ChF'_{hi}(w'_i)$ Характеристической Функции ChF'_{hi} для w'_i;

Тернарное отношение $I \subseteq G \times M \times V$ задает частичное отображение G на $V : m(g) = v$, ставит в соответствие каждой ХФ ее значение для заданного w_i.

Решетка Формальных Понятий для Лексического Значения.

Определение 3. *Под Формальным Понятием (ФП) для (5) понимается пара* $(X, Y) : X \subseteq G$, $Y \subseteq M \times V$, $X = Y'$, $Y = X'$, *причем*

$$X' = \{(m, v) : m \in M, v \in V | \forall g \in X : m(g) = v\},$$

$$Y' = \{g \in G | \forall (m, v) \in Y : m(g) = v\}$$

Определение 4. *ФП* (X_1, Y_1) *является подпонятием для ФП* (X_2, Y_2), *если* $X_1 \subseteq X_2$, *а* $Y_2 \subseteq Y_1 : (X_1, Y_1) \leq (X_2, Y_2)$. *При этом* (X_2, Y_2) *называют суперпонятием для ФП* (X_1, Y_1), *а отношение* $\leq$ *— отношением порядка для ФП.*

Определение 5. *Множество* $\Re(G, M, V, I)$ *всех ФП контекста (5) вместе с отношением* $\leq$ *называется решеткой Формальных Понятий.*

Определение 6. *Пусть* $N \subset \Re(G, M, V, I)$. *ФП* (X, Y) *называется Наименьшим Общим Суперпонятием (НОСП) для* N, *если* $(X_i, Y_i) \leq (X, Y)$ *для* $\forall (X_i, Y_i) \in N$ *и* $\nexists (X_1, Y_1) \in \Re(G, M, V, I) \setminus N : (X_1, Y_1) \leq (X, Y)$ *и* $(X_i, Y_i) \leq (X_1, Y_1)$ *для* $\forall (X_i, Y_i) \in N$. *Аналогично определяется Наибольшее Общее Подпонятие (НОПП) для* N.

Определение 7. *Под областью в решетке Формальных Понятий для (5) понимается набор ФП, связанных отношением* $\leq$ *с одним НОПП и/или одним НОСП.*

Замечание. *В настоящей работе для областей вводится требование единственности как НОПП, так и НОСП.*

Пример построения формального контекста для независимых толкований заданного Лексического Значения выделением множества Характеристических Функций.

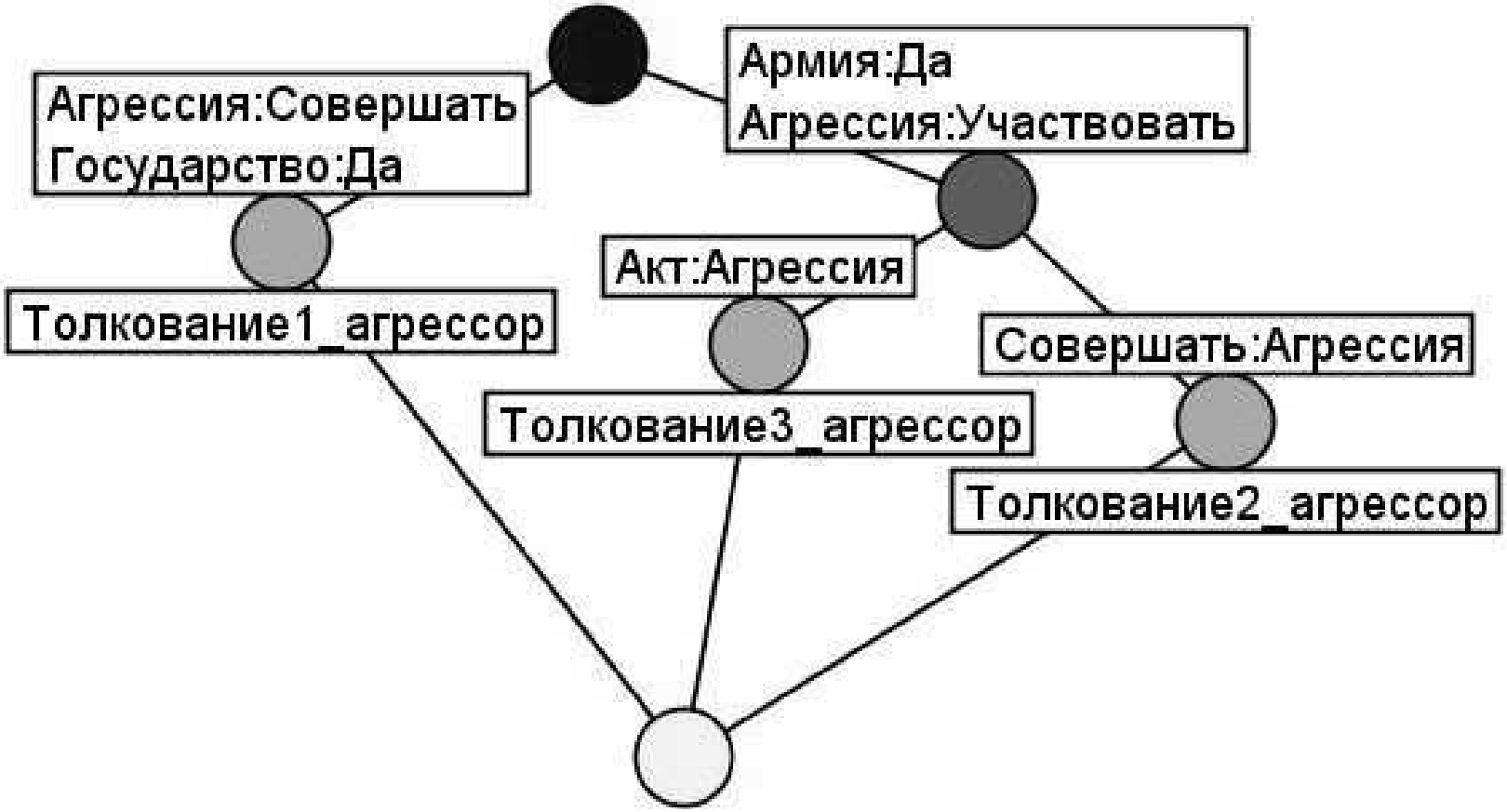

Рис. 3. Формализованные толкования для ЛЗ «агрессор»

Обобщение утверждений независимых теорий ЛЗ.

Утверждение 1. *Утверждения* (R_n, C, L_1^M) *и* (R_n, C, L_2^M) *вида (3) могут быть представлены одним утверждением :*

$$(R_n, C, \{(\text{«или»}, L_3^M)\}),$$

если наборы ФП, полученные на основе L_1^M, L_2^M *и* L_3^M *образуют области* $\Re(G_1, M_1, V_1, I)$, $\Re(G_2, M_2, V_1, I)$ *и, соответственно,* $\Re(G_3, M_3, V_1, I)$ *с НОСП, которое имеет* R_n *в качестве значения признака. При этом :*

$$G_1 = \{(w_i', L_1^M)\}, G_2 = \{(w_i'', L_2^M)\}, M_1 \neq M_2, M_3 = M_1 \cup M_2,$$
$$\Re(G_3, M_3, V_1, I) = \Re(G_1, M_1, V_1, I) \cup \Re(G_2, M_2, V_1, I).$$

Утверждение 2. *Утверждения* (R_n, C, L_1^M) *и* (R_n, C, L_2^M) *вида (3) могут быть представлены одним утверждением :*

$$R_n, C, \{(\text{«и»}, L_3^M)\},$$

если на основе L_1^M, L_2^M *и* L_3^M *определяются ФП* (X, Y_1), (X, Y_2) *и* (X, Y_3) *:* $Y_3 = Y_1 \cup Y_2$.

Замечание. *Согласно Определению 2, внешне различные описания теорий (1) одного и того же ЛЗ задают единое множество Характеристических Функций. Следовательно, мощность n множества ХФ для заданного ЛЗ не зависит от количества k обобщаемых теорий. Вычислительная сложность процесса обобщения теорий для заданного ЛЗ составляет* $O\left(\frac{n}{k}\right)^k$. *Поскольку* $k \in [1, \ldots, n]$, *то* $O\left(\frac{n}{k}\right)^k = n$ *при* $k = 1$ *и* $O\left(\frac{n}{k}\right)^k = 1$ *при* $k = n$.

Пример обобщения независимых теорий заданного ЛЗ.

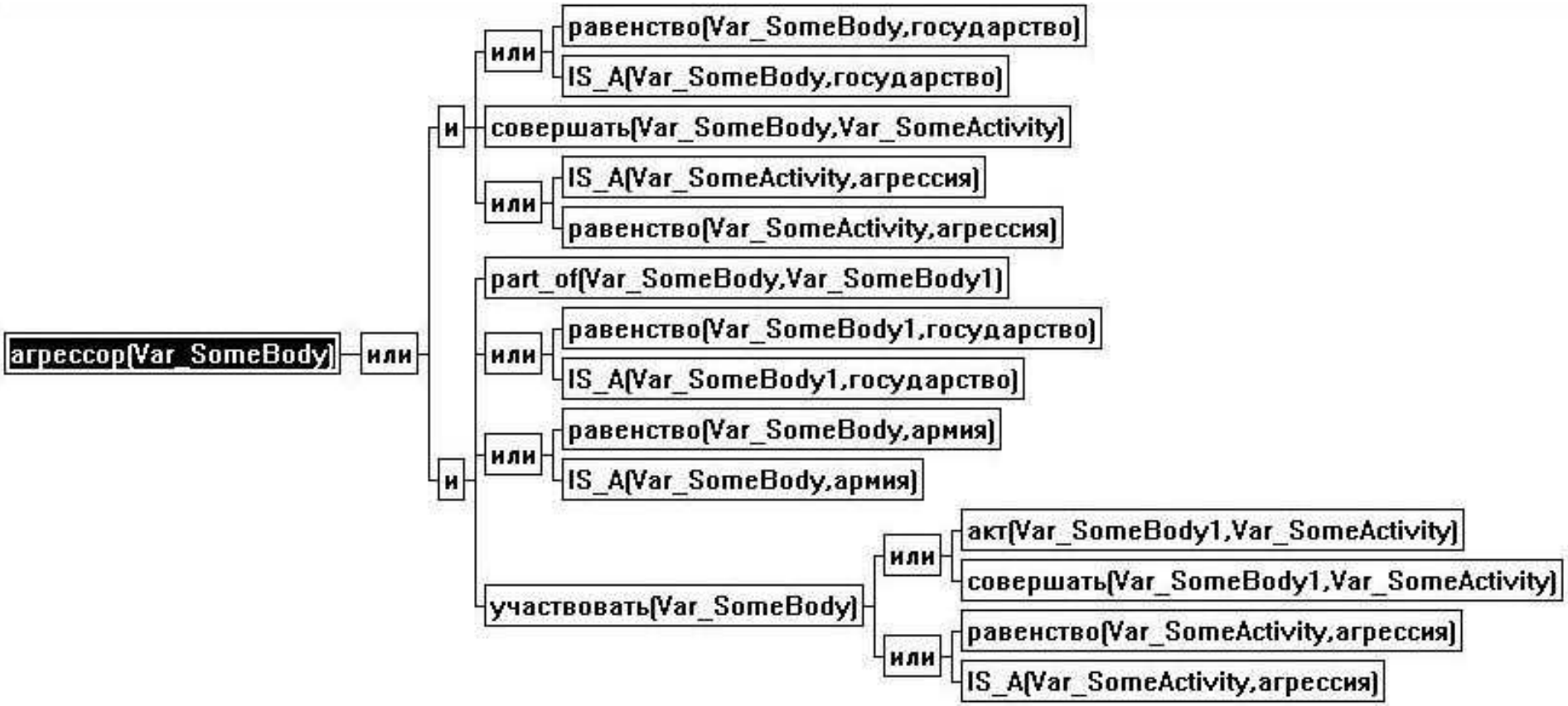

Рис. 4. Обобщенная теория ЛЗ «агрессор»

Смысловое отношение в рамках Расщепленного Значения.

Пусть :

Π^R — множество правил синонимических преобразований ЕЯ-высказываний;

L^{Π} — множество пар ЕЯ-высказываний, между которыми возможно установление синонимии (относительно Π^R);

$T = \{T_1, T_2\} : T \in L^{\Pi}$.

Утверждение 3. *Смысловое отношение F, значимое для формирования $r(\pi)$, между некоторым словом $w_1 \in T_1$ и его лексическим коррелятом $w_2 \in T_2$, входящим в РЗ, будет иметь место тогда, когда*

$$L_1^M = L_{11}^M \cup \left\{(F, C, L_{22}^M)\right\} \cup L_{12}^M,$$

$$L_2^M = L_{11}^M \cup L_{22}^M \cup L_{12}^M,$$

$$L_{11}^M \cap L_{22}^M = \emptyset, L_{11}^M \cap L_{12}^M = \emptyset, L_{12}^M \cap L_{22}^M = \emptyset,$$

где L_1^M — набор утверждений теории ЛЗ для w_1, а L_2^M — для w_2.

Пример — теории ЛЗ «эксперимент» и «экспериментировать», Рис. 5.

Пример использования Лексической Функции в качестве названия Смыслового Отношения в теории Лексического Значения.

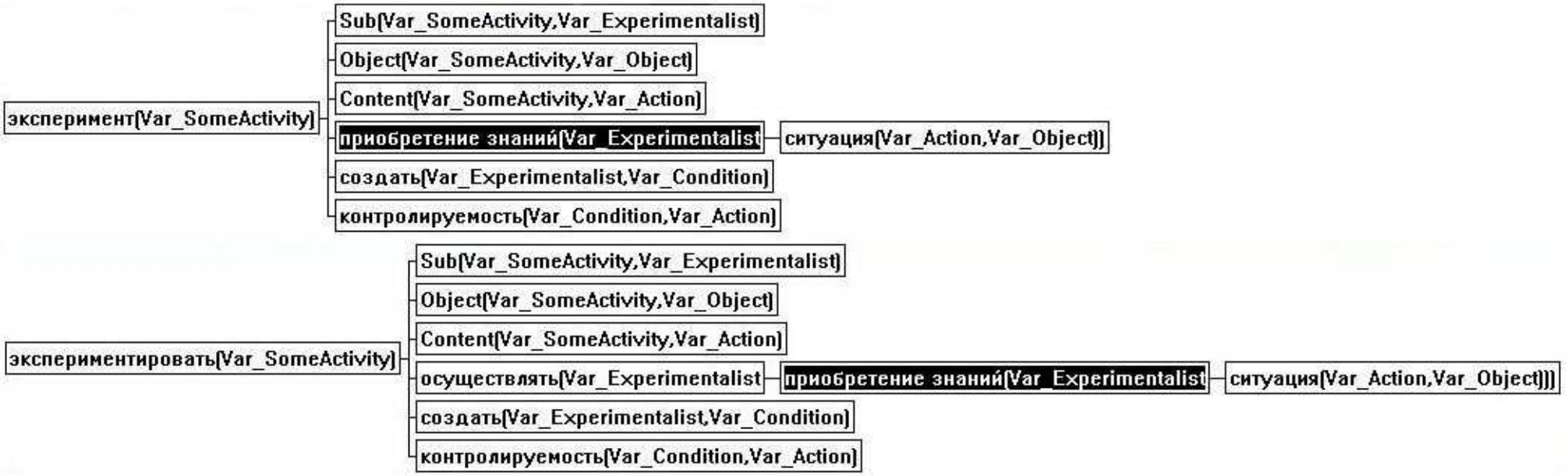

Рис. 5. Теории ЛЗ «эксперимент» и «экспериментировать»

Формальный контекст для совокупности слов-аргументов Лексической Функции-параметра.

$$K^{LF} = (G^{LF}, M^{LF}, I^{LF}) \tag{6}$$

Здесь :

множество объектов G^{LF} есть множество ключевых слов-аргументов заданной Лексической Функции;

Множество формальных признаков M^{LF} есть множество слов-значений заданной Лексической Функции для слов из множества G^{LF};

Бинарное отношение $I^{LF} \subseteq G^{LF} \times M^{LF}$ задает частичное отображение G^{LF} на M^{LF} и ставит в соответствие каждому ключевому слову $C_0 \in G^{LF}$, определяющему ситуацию СЭ, множество значений заданной Лексической Функции.

Модель системы слов-аргументов заданной Лексической Функции-параметра.

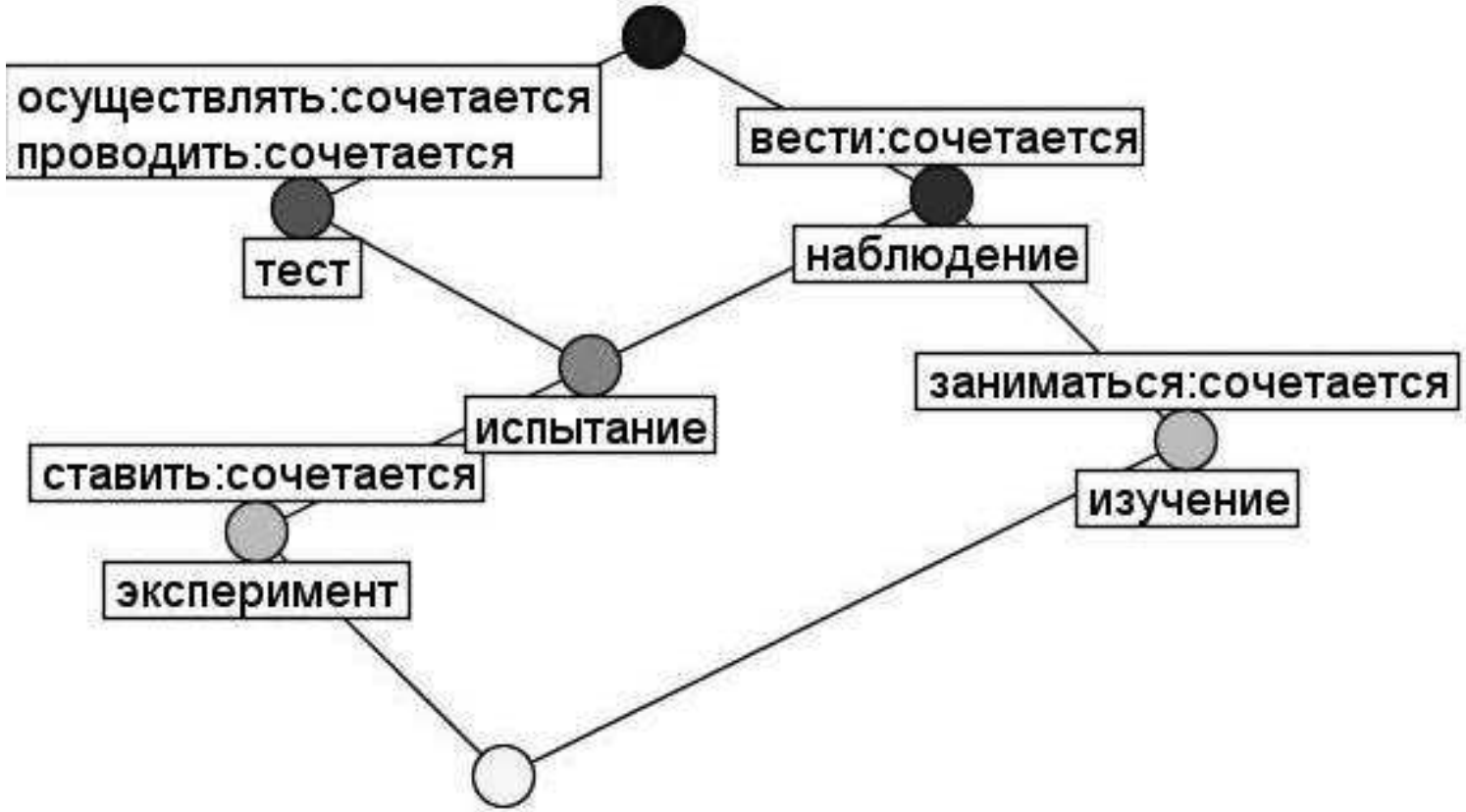

Рис. 6. Слова-аргументы Лексической Функции $Oper_1$ из верхней окрестности для Лексического Значения «эксперимент»

Отношение порядка для Лексических Значений предикатных слов.

Пусть для $\forall w_i$ (1) мы имеем описание ее СК C_i :

$$S_i^F = (C_i, L_i^{SF}, D_i, D_i') : \tag{7}$$

- списком L_i^{SF} дескрипторов Семантических Характеристик (СХ) слова w_i, обозначающего сущность C_i, в последовательности «более общая СХ — более специфическая СХ»;
- дескрипторами таксономической категории D_i и ее подкласса D_i' для w_i.

Предположим также, что w_i обозначает некоторую ситуацию. При этом для w_i имеется описание характеризованного ролевого состава :

$$C^A = (C_i, L_i^R), \tag{8}$$

где $\forall A_{ti} \in L_i^R$ включает название R_{ti} роли плюс список L_{ti}^C возможных Семантических Классов актанта :

$$A_{ti} = (R_{ti}, L_{ti}^C) \tag{9}$$

Утверждение 4. *ЛЗ Семантического Класса C_1 :*

$$C_1^A = (C_1, L_1^R)$$

следует считать суперпонятием для ЛЗ Семантического Класса C_2 :

$$C_2^A = (C_2, L_2^R)$$

при условии, что для $\forall R_t : (R_t, L_{t2}^C) \in L_2^R \; \exists (R_t, L_{t1}^C) \in L_1^R$: каждому $C_{at1} \in L_{t1}^C$ можно поставить в соответствие $C_{at2} \in L_{t2}^C$: либо $C_{at2} = C_{at1}$, либо C_{at2} и C_{at1} связаны отношением IS_A.

Отношение порядка для случая зависимости между Семантическими Характеристиками предикатных слов.

Утверждение 5. *ЛЗ* w_i *Семантического Класса* C_i *(7) :*

$$S_i^F = (C_i, L_i^{SF}, D, D')$$

будет считаться суперпонятием для ЛЗ w_m *Семантического Класса* C_m *:*

$$S_m^F = (C_m, L_m^{SF}, D, D'), w_i \neq w_m,$$

если в дополнение к определенным Утверждением 4 условиям при отсутствии для $A_{ai} = (R_{ai}, L_{ai}^C) : A_{ai} \in L_i^R$*, (8,9), актанта подпонятия с показанным в Утверждении 4 соответствием набора возможных СК* $\exists A_{bm} = (R_{bm}, L_{bm}^C) : A_{bm} \in L_m^R$*, отвечающий нижеследующему требованию. При наличии*

$$S_{qai}^F = (C_{qai}, L_{qai}^{SF}, D_{qai}, D'_{qai})$$

для $\forall C_{qai} \in L_{ai}^C$ *и, соответственно,*

$$S_{sbm}^F = (C_{sbm}, L_{sbm}^{SF}, D_{sbm}, D'_{sbm})$$

для $\forall C_{sbm} \in L_{bm}^C$ *наряду с вхождением в* L_{sbm}^{SF} *СХ из списка* L_{qai}^{SF} *некоторым СХ* $SF_{pqai} \in L_{qai}^{SF}$ *ставятся в соответствие теории (1) :*

$$Lm_{pqai} = (SF_{pqai}, L_{pqai}^M),$$

причем $\exists L_{sbm}^{SF\prime} \subset L_{sbm}^{SF} : \forall SF_{osbm} \in L_{sbm}^{SF\prime}$ *является в составе* L_{pqai}^M *либо одним из аргументов структуры (2), либо первым аргументом структуры (3).*

Пример : валентность аспекта у ЛЗ «испытание» и валентность содержания у ЛЗ «тест».

Имеем :

w_i =«тест», w_m =«испытание»

S^F_{qai} = («ситуация»,[«SITUAT»],«LABL»,«SIT»),

S^F_{sbm} = («свойство»,[«ATTR»],«ASP»,«Не определена»).

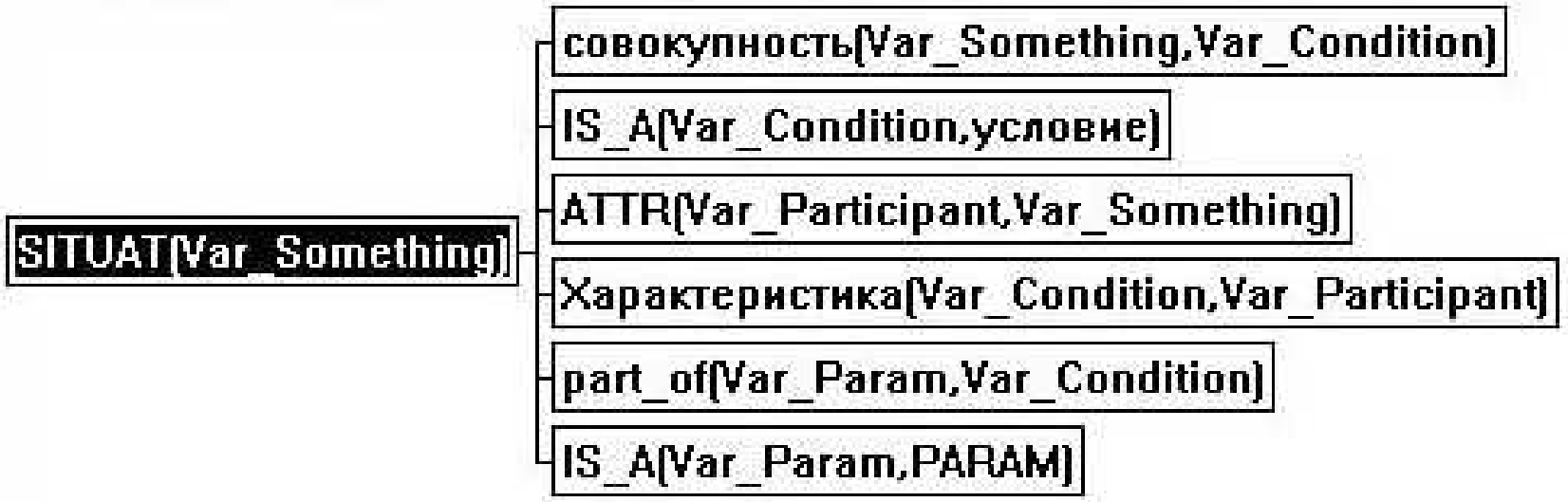

Рис. 7. Теория сорта «SITUAT»

Расширенное отношение порядка для слов-аргументов ЛФ $Oper_1$.

Рис. 8. СК слов окрестности ЛЗ «эксперимент»

Таблица 1. СК для слов окрестности ЛЗ «эксперимент»

Слово	Семантичекий класс
эксперимент	получение знаний об объекте или явлении при контролируемых условиях
испытание	действие с целью получения знаний при сопутствующем наблюдении
изучение	получение знаний
тест	действие с целью получения знаний
наблюдение	целенаправленное восприятие

Критерий адекватности формирования прецедента.

Пусть для ЛЗ $w_1 \in T_1$ и $w_2 \in T_2 : T = \{T_1, T_2\}$, $T \in L^{\Pi}$ имеются описания теорий $Lm(w_1)$ и $Lm(w_2)$ в соответствии с (1).

Кроме того, имеем : $W^S : \forall w_i \in W^S$ обозначает некоторую ситуацию и $w_i \in T_j$, $j \in \{1, 2\}$. Для $\forall w_i \in W^S$ имеем описание характеризованного ролевого состава в форме (8).

Утверждение 6. *Будем считать, что* $Lm(w_1)$ *и* $Lm(w_2)$ *:* $w_1 \in W^S$, $w_2 \in W^S$ *адекватно задают* $r(\pi)$ *в соответствии с Определением 1, если :*

- *на множестве* W^S *может быть определено отношение порядка в соответствии с условиями в Утверждениях 4 и 5;*
- *между* w_1 *и* w_2 *существует смысловое отношение* F *в соответствии с условиями, задаваемыми Утверждением 3;*
- F *в составе формального контекста (6) принадлежит множеству формальных признаков того ЛЗ* w_{Sup}*, которое является в соответствии с Определением 6 Наименьшим Общим Суперпонятием для множества* N^A *слов верхней окрестности ЛЗ* w_2*,* $N^A \subset \Re(G^A, M^A, V^A, I^A)$ *:*

  $G^A \supset W^S$ *— множество ЛЗ предикатных слов,*

  M^A *— множество возможных ролевых ориентаций* R_{ti} *актантов (9) для обозначаемых предикатными словами* $w_m \in G^A$ *ситуаций,*

  V^A *— множество всех множеств* L^C_{ti} *СК слов, способных замещать некоторую валентность* R_{ti} *(9) предикатного слова* $w_m \in G^A$*,*

  $I^A \subseteq G^A \times M^A \times V^A$.

Требования к РЗ с w_{Sup} определяются аналогично.

Выводы.

— Описание смысла слова набором Характеристических Функций производится в шкале наименований. При обобщении утверждений независимых теорий одного и того же Лексического Значения посредством отношения «или» не учитывается статистическая значимость каждого признака. Значения Характеристических Функций, которые задаются объединяемыми утверждениями, полагаются равновероятными.

— Перспективным направлением дальнейших исследований является введение в рассмотрение распределений возможных значений Характеристических Функций. Это позволит вычислять меру близости между предикатами, описывающими Лексическое Значение слова в рамках $r(\pi)$ и тем самым сократить объем обучающей выборки при формировании $r(\pi)$.

— Задействование Характеристических Функций при описании смысла слова и их выводимость из теории его Лексического Значения позволяет в перспективе ввести в рассмотрение родовидовые зависимости между теориями при описании $r(\pi)$ для ситуаций СЭ на основе неточных синонимов, конверсивов и дериватов. Это позволит фиксировать различия в актантной структуре этих слов.