Машинное понимание текстов в общей задаче распознавания образов.

Михайлов Д. В., Емельянов Г. М.

Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

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

Предмет исследования.

Семантическая близость текстов предметно-ориентированного подмножества Естественного Языка (ЕЯ).

Задачи.

1) Формальное определение смыслового эталона.
2) Разработка структуры базы данных эталонов для анализа близости текстов.
3) Введение меры семантической близости на основе знаний о ситуациях смысловой эквивалентности для подмножества ЕЯ.

Анализ формальных понятий и ситуации языкового употребления.

Представим языковой контекст Ситуации Языкового Употребления (СЯУ) посредством Формального Контекста (ФК):

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

где множество объектов G составляют основы слов, синтаксически подчиненных другим словам. Множество признаков M включает подмножества, обозначаемые далее посредством M с соответствующим нижним индексом и содержащие:

- указания на основу синтаксически главного слова (M_1);
- указания на флексию главного слова (M_2);
- связи «основа–флексия» для синтаксически главного слова (M_3);
- сочетания флексий зависимого и главного слова (M_4). При этом после флексии главного слова через двоеточие указывается предлог (если такой имеется) для связи с зависимым;
- указания на флексию зависимого слова (M_5).

Определение 1. *Пара множеств (A, B), называемых объемом и содержанием, образуют формальное понятие (ФП), если имеют место отображения:*

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

где $A' = B,\ B' = A,\ I \subseteq G \times M$ ставит в соответствие объектам их признаки.

Определение 2. *Множество $\Re(G, M, I)$ всех ФП формального контекста вместе с отношением порядка называется решеткой формальных понятий.*

Определение 3. *ФП вида (g'', g') называется объектным ФП, аналогично ФП вида (m', m'') считается признаковым ФП, где $g \in G, m \in M$.*

Пусть $K^E = \left(G^E, M^E, I^E\right)$ есть ФК СЯУ S_1 для корректного ЕЯ-описания некоторого факта, $K^X = \left(G^X, M^X, I^X\right)$ — ФК произвольной СЯУ S_2.

Введем обозначения для символьных констант: p_{fl} — «флексия:», p_{bs} — «главное-основа:», p_b — «основа:», а для операции конкатенации — символ $\odot$.

Расщепленные предикатные значения.

Теорема 1. *Пусть* $\{m_1, m_2, m_3\} \subset M_1$. *Если считать признаки* m_1, m_2 *и* m_3 *взаимно различными, то* m_1 *соответствует указанию на основу главного,* m_2 — *зависимого слова Расщепленного Предикатного Значения (РПЗ),* m_3 — *однословного смыслового эквивалента этого РПЗ при выполнении трех условий:*

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) $\exists\, \{g_2, g_3\} \subset G$, *при этом объекты* g_1, g_2 *и* g_3 *взаимно различны, а*

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

3) не существует других троек объектов, для которых признак m_3 *занимал бы место либо признака* m_1, *либо признака* m_2 *в вышеуказанных соотношениях.*

Замечание 1. *После удаления информации РПЗ формальный контекст СЯУ отражает классы отношений, которые определяются исключительно ролями объектов-участников ситуации по отношению к ней самой.*

Замечание 2. *Слова-синонимы могут обозначать понятия с различной степенью абстракции. Указанная степень тем более, чем больше количество СЯУ, относительно которых понятие фигурирует в некоторой фиксированной роли.*

Формирование тезауруса на основе совокупности СЯУ.

Рассмотрим модель тезауруса в виде формального контекста:

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

где G^H состоит из пометок отдельных СЯУ. Множество M^H содержит элементы множеств признаков ФК всех $g^H \in G^H$. Кроме того, в составе M^H выделяются:

- M_6 — множество указаний на объекты ФК отдельных $g^H \in G^H$;
- M_7 — множество связей «основа–флексия» для синтаксически зависимого слова;
- M_8 — множество сочетаний основ зависимого и главного слова.

Обозначим объединение множеств M_6, M_7, M_8, M_4^E, M_4^X, M_5^E и M_5^X, как M^U.

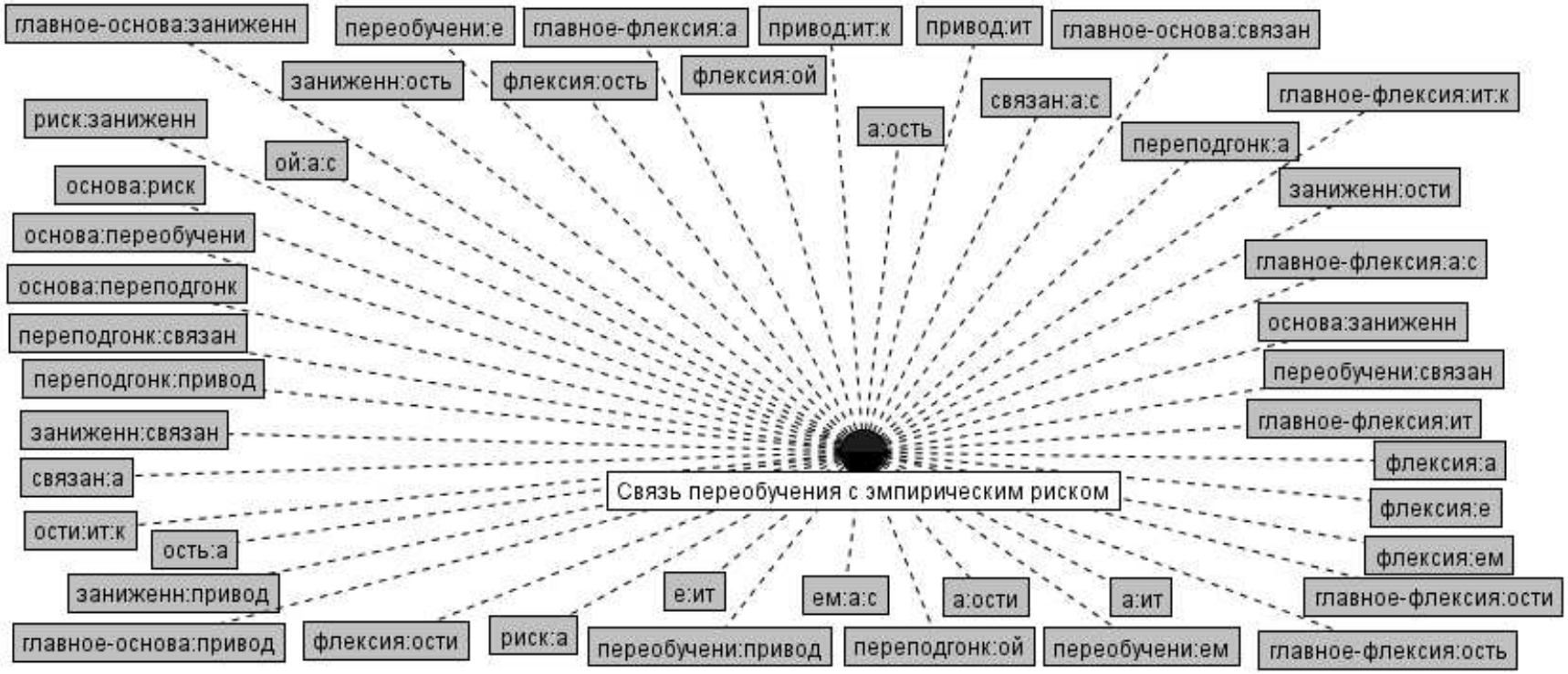

Рис. 1. Объект $g^H \in G^H$ для формального контекста отдельной СЯУ.

Определение 4. *Будем считать, что S_1 и S_2 связаны отношением схожести, если каждому объекту $g^X \in G^X$ соответствует такой объект $g^E \in G^E$, что выполняется одно из условий:*

1) $g^X = g^E$ и любой признак $m^E \in M^E$ объекта g^E будет относиться и к объекту g^X.

2) $g^X = g^E$, при этом Условие 1) не выполняется, но существует объект $g^H \in G^H$, обладающий признаком $m_1^H \in M_6 \colon m_1^H = p_b \odot g^E$ при обязательном выполнении следующих условий:

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

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

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

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

Кроме того, для $\forall\, m^H \in (M^H \backslash M^U)$ верно:

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

3) $g^X \neq g^E$, но существует объект $g^H \in G^H$, обладающий признаками $m_1^H \in M_6 :\ m_1^H = p_b \odot g^E$ и $m_2^H \in M_6 :\ m_2^H = p_b \odot g^X$, при этом для любого признака $m^H \in (M^H \backslash M^U)$ справедливо:

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

4) $g^X \neq g^E$, но $\exists\left(g_1^H \in G^H, m_1^H \in M_6\right) : I^H\left(g_1^H, m_1^H\right) = \text{true}$, $m_1^H = p_b \odot g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ верно $\left(I^H\left(g_1^H, m_1^H\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^H\left(g_1^H, m^E\right)$. При этом имеются признаки $m_2^H \in M_6$ и $m^X \in \left(M_1^X \cup M_2^X \cup M_3^X\right)$, для которых $\left(I^H\left(g_1^H, m_2^H\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^H\left(g_1^H, m^X\right)$, где $m_2^H = p_b \odot g^{X_1}$, $g^{X_1} \neq g^X$, а пара $\left(g^{X_1}, g^E\right)$ отвечает Условию 3) настоящего Определения при генерации формального контекста для СЯУ g_1^H. В то же время существует объект $g_2^H \in G^H$, относительно которого пара $\left(g^X, g^{X_1}\right)$ также будет отвечать Условию 3) настоящего Определения. Генерируемый при этом формальный контекст для СЯУ g_2^H будем обозначать далее как K^{X_1}. По аналогии с K^E и K^X, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.

Мера близости ситуаций языкового употребления.

Мера близости СЯУ S_1 и S_2 относительно ФК $K^E = (G^E, M^E, I^E)$ и $K^X = (G^X, M^X, I^X)$, из которых удалена информация РПЗ, определяется по формуле:

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

где $n = |G^X|$, а spc_k есть мера мера близости объектов в паре (g_k^X, g^E). Значение spc_k:
- равно 1.0, если выполнено *Условие 1) Определения 4*;
- вычисляется по формуле:

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

если выполнено *Условие 2), 3), либо 4).*

Если $\exists\, g^X \in G^X$, для которого нет выполнимых условий *Определения 4*, то $spc(S_1, S_2) = 0$.

В случае истинности любого из *Условий 2)–4) Определения 4* значение $D_c = 2$.

При выполнении *Условия 2)* либо *3)* число $path_C = 4$, а в множество B^C войдут признаки $m^H \in (M^H \backslash M^U)$, для которых справедливо либо соотношение (3) (при выполнении *Условия 2)*), либо соотношение (4) (при выполнении *Условия 3)*). При этом

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

Выполнимость *Условия 4)* обычно проверяется в несколько итераций. В ходе каждой итерации число признаков, не являющихся общими для g_k^X и g^{X_1}, всегда меньше, чем в предыдущей. Начальное значение $path_C = 4$ и с каждым шагом возрастает на 1. При истинном *Условии 4)*

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

где $\left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right) \subset M^{X_1}$. Множество B^C здесь есть пересечение B_1 и B_2.

Таблица 1. Исходные данные для построения фрагмента тезауруса.

№п/п	1				2	3		4	
основа	флективная часть + предлог								
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

Пример исходных данных для построения формальных контекстов сравниваемых ситуаций языкового употребления.

Таблица 2. Описание факта связи между переобучением и эмпирическим риском.

ЕЯ-описание	эталонное				анализируемые			
вариант	1	2	3	4	1	2	3	4
основа	флективная часть + предлог							
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

Примечание. *Вариант № 4 анализируемого описания рассматриваемого факта — неправильный: «Заниженность средней ошибки на обучающей выборке приводит к эмпирическому риску».*

Результат : значения близости эталону для анализируемых вариантов описания заданного факта предметной области.

Таблица 3. Сравнение вариантов ЕЯ-описания связи между переобучением и эмпирическим риском.

Вариант	$spc(S_1, S_2)$	$\left\|B^C\right\|$	$\left\|B_1 \backslash B^C\right\|$	$\left\|B_2 \backslash B^C\right\|$
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	—	—	—

Выводы.

- Основной *результат* работы — *метод анализа близости ситуаций языкового употребления при их независимом порождении.* Использованием унифицируемого теоретико-решеточного представления сравниваемых ЕЯ-высказываний и тезаурусной информации достигается максимальная простота пополнения тезауруса и эффективность его использования при анализе близости текстов.
- Предложенная *модель тезауруса* может служить основой базы данных смысловых эталонов для текстов по заданной предметной области. Иерархическое представление информации в решетке формальных понятий позволяет уменьшить как размер базы эталонов, так и время поиска в ней.
- Модель тезауруса в виде решетки формальных понятий обеспечивает сжатие информации (в первую очередь) за счет тех предикатных слов, которые обозначают ситуации, сходные в той или иной мере по составу участников и характеру выполняемых ими действий, а также за счет абстрактной лексики. Степень сжатия информации зависит от релевантности заданной предметной области каждого из представленных в решетке ЕЯ-описаний отдельных фактов.
- Отдельного прикладного исследования заслуживают количественные оценки полноты охвата языкового описания предметных знаний в решетке тезауруса.