Семантическая схожесть текстов в задаче автоматизированного контроля знаний*

Михайлов Д. В., Емельянов Г. М.
Dmitry.Mikhaylov@novsu.ru
Великий Новгород, ГОУ ВПО «Новгородский государственный университет имени Ярослава Мудрого»

В данной статье описывается применение методов Анализа Формальных Понятий для интерпретации результатов тестов открытой формы в системах контроля знаний. Вводится мера семантической схожести между ответом обучаемого и вариантом правильного ответа, задаваемого разработчиком теста.

Semantic affinity of texts in a problem of computer-aided testing of knowledge*

Mikhaylov D. V., Emelyanov G. M.
Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russia

The approach offered is to apply the methods of Formal Concept Analysis for interpretation of the results of open form's test task in system of computer-aided testing of knowledge. The measure of semantic affinity between trainee's answer and correct answer's variant given by test's developer, is introduced.

Тестовое задание открытой формы [4] в системе контроля знаний предполагает ответ обучаемого в виде одного или нескольких предложений Естественного Языка (ЕЯ). Как правило, разработчик теста формулирует свой вариант правильного ответа на основе собственных знаний по заданной Предметной Области (ПО). Традиционно интерпретация ответа обучаемого здесь заключается в простом поиске среди «правильных» вариантов [3].

Актуальная при этом *проблема* адекватности результатов тестирования обусловлена как лексико-функциональной синонимией [2] в ответах, так и недостаточным задействием предметных знаний.

В настоящей работе рассматривается анализ степени близости ответа обучаемого заданному эталону с помощью тезауруса, формируемого на основе множеств вариантов правильных ответов по совокупности тестов заданной тематики

Постановка задачи

Положим отдельный факт некоторой ПО описанным множеством Семантически Эквивалентных (СЭ) ЕЯ-фраз, которые определяют Ситуацию Языкового Употребления (СЯУ, [2]). Представим языковой контекст, фиксируемый СЯУ, посредством тройки вида

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

именуемой Формальным Контекстом (ФК, [1]).

При этом множество объектов G составляют основы слов, синтаксически подчиненных другим словам из СЭ-фраз, задающих СЯУ. Множество признаков M включает в себя 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 главного слова с зависимым;
- указания на флексию зависимого слова (M_5).

Рассмотрим совокупность СЯУ для известных фактов заданной предметной области как основу формирования тезауруса. Будем рассматривать модель тезауруса в виде формального контекста:

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

где множество G^H состоит из символьных пометок отдельных СЯУ. Множество M^H содержит элементы множеств признаков ФК вида (1) всех $g^H \in G^H$. Кроме того, в составе M^H выделяются:

- множество указаний на основы слов, синтаксически подчиненных другим словам в ЕЯ-описаниях ситуаций $g^H \in G^H$. Фактически данное множество, обозначаемое далее как M_6, содержит указания на объекты ФК вида (1), генерируемых для элементов множества G^H;
- множество связей «основа–флексия» для синтаксически зависимого слова, M_7;
- множество сочетаний основ зависимого и главного слова, M_8.

Отношения $I \subseteq G \times M$ и $I^H \subseteq G^H \times M^H$ ставят в соответствие объектам множеств G и G^H их признаки из множеств M и M^H, соответственно. При этом ФК обоих видов могут быть сформированы по результатам синтаксического анали-

Работа выполнена при поддержке УНИК НовГУ и РФФИ, проект № 10-01-00146.

за соответствующих ЕЯ-фраз. В настоящей работе мы рассмотрим *совместное использование моделей* (1) *и* (2) для вычисления *меры схожести СЯУ* при интерпретации тестового задания открытой формы.

Методы решения

Пусть $K^E = \left(G^E, M^E, I^E\right)$ есть ФК вида (1) для СЯУ S_1, соответствующей сформулированному разработчиком теста варианту правильного ответа. Введем также в рассмотрение аналогичную структуру $K^X = \left(G^X, M^X, I^X\right)$ для СЯУ S_2, задаваемой ответом обучаемого.

Обозначим множество, получаемое объединением множеств M_6, M_7, M_8, M_4^E, M_4^X, M_5^E и M_5^X, как M^U. Введем также обозначения для используемых далее символьных констант: p_{fl} — для «флексия:», p_{bs} — для «главное-основа:», p_{b} — для «основа:», а для операции конкатенации — символ $\odot$.

Определение 1. *Будем считать, что СЯУ 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 следующих 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 $m_1^H \in M_6 \colon m_1^H = p_{\mathrm{b}} \odot g^E$ при обязательном выполнении следующих условий:*

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

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

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

при этом $I^E\left(g^E, m_{\mathrm{bs}}^E\right) \rightarrow I^H\left(g^H, m_{18}^H\right)$;

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

при этом $I^X\left(g^E, m_{\mathrm{bs}}^X\right) \rightarrow I^H\left(g^H, m_{28}^H\right)$.

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

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

В содержательном плане Условие 2) настоящего Определения описывает случай наличия синонимов среди слов, синтаксически главных по отношению к словам со сходными основами. При этом основы g^X и g^E не омонимичны, поскольку было бы нарушено требование разделения ими признаков главного слова.

3) *$g^X \neq g^E$, но существует объект $g^H \in G^H$, обладающий признаками $m_1^H \in M_6 \colon m_1^H = p_{\mathrm{b}} \odot g^E$ и $m_2^H \in M_6 \colon m_2^H = p_{\mathrm{b}} \odot g^X$, при этом для любого признака $m^H \in \left(M^H \backslash M^U\right)$ справедливо:*

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

4) *$g^X \neq g^E$, но существует объект $g_1^H \in G^H$, обладающий признаком $m_1^H \in M_6 \colon m_1^H = p_{\mathrm{b}} \odot g^E$, а для $\forall\, m^E \in \left(M_4^E \cup M_5^E\right)$ справедливо:*

$$\left(I^H\left(g_1^H, m_1^H\right) \wedge I^E\left(g^E, m^E\right)\right) \rightarrow I^H\left(g_1^H, m^E\right).$$

При этом существуют 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 верно:

$$\left(I^H\left(g_1^H, m_2^H\right) \wedge I^X\left(g^X, m^X\right)\right) \rightarrow I^H\left(g_1^H, m^X\right),$$

где $m_2^H = p_{\mathrm{b}} \odot g^{X_1}$, $g^{X_1} \neq g^X$, а пара $\left(g^{X_1}, g^E\right)$ отвечает Условию 3) настоящего Определения при генерации формального контекста вида (1) *для объекта 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 вида* (1) *для объекта g_2^H будем обозначать далее как K^{X_1}. По аналогии с K^E и K^X, $K^{X_1} = \left(G^{X_1}, M^{X_1}, I^{X_1}\right)$.*

Замечание 1. Анализ схожести S_1 и S_2 включает сравнение последовательностей двух и более соподчиненных слов. Пример: *«средняя ошибка на обучающей выборке» ⇔ «эмпирический риск»*. Выполнимость условий *Определения 1* здесь анализируется только для главных слов (в примере это *«ошибка»* и *«риск»*). Сами последовательности считаются взаимно заменяемыми, если возможно их построение по ФК (2) на наборе признаков с префиксом p_{bs} для одной и той же СЯУ. При этом главные слова последовательностей должны быть одинаково подчинены одному и тому же слову, что проверяется по сочетанию флексий.

Как следует из *Определения 1*, оценка схожести СЯУ производится относительно классов Формальных Понятий (ФП) для формальных контекстов K^E и K^X, соответственно. Каждое ФП формального контекста вида (1) есть пара множеств (A, B), называемых объемом и содержанием [1], причем существуют отображения:

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

Множество всех ФП формального контекста вместе с отношением порядка называют *решеткой ФП*.

Классы понятий в решетке для формального контекста вида (1) различаются степенью абстракции, которая зависит от частоты употребления главных слов рассматриваемых сочетаний в различных синтаксических контекстах.

Схожесть СЯУ оценивается относительно классов формальных понятий одного уровня абстракции, соответствующих подчинению существительных тем словам, которые называют ситуацию, но не входят в Расщепленные Предикатные Значения (РПЗ). Каждое РПЗ есть совокупность вспомогательного глагола (связки) и некоторого существительного, называющего ситуацию. Выделение объектов и признаков РПЗ из структуры (1) основано на следующей теореме.

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 РПЗ, а* m_3 *— указанию на основу однословного смыслового эквивалента этого РПЗ при выполнении трех условий:*

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

После удаления информации РПЗ формальный контекст (1) отражает классы отношений, определяемых ролями участников описываемой ситуации действительности по отношению к ней самой.

Мера схожести S_1 и S_2 относительно формальных контекстов K^E и K^X, из которых удалена информация РПЗ, определяется по формуле:

$$\text{spc}(S_1, S_2) = \frac{\sum_{k=1}^{n} \text{spc}_k}{n}, \qquad (5)$$

где $n = \left|G^X\right|$, а spc_k есть мера мера схожести объектов в паре $\left(g_k^X, g^E\right)$. В зависимости от выполнимости условий *Определения 1*, значение spc_k:

— равно 1.0, если выполнено *Условие 1)*;
— вычисляется по формуле:

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

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

Во втором случае мы имеем дело с гипотетической решеткой ФП (обозначим ее как Re^{XE}), в которой объемы объектных ФП (ФП с одним объектом в объеме) есть $\left\{g_k^X\right\}$ и $\left\{g^E\right\}$ (при выполнении *Условия 2)* или *3)*), либо $\left\{g_k^X\right\}$, $\left\{g^E\right\}$ и $\left\{g^{X_1}\right\}$ (при выполнении *Условия 4)*). Значение D_c равно количеству сравнимых ФП, составляющих цепочку с вершинным ФП решетки Re^{XE} в качестве максимального ФП и Наименьшим Общим Суперпонятием (НОСП) для объектных ФП решетки Re^{XE} — в качестве минимального ФП. Множество B^C есть содержание этого НОСП, а число path_C равно минимальному количеству ФП в цепочке, которой принадлежит вершинное ФП, наименьшее ФП решетки Re^{XE} и ФП с содержанием B^C.

В случае истинности любого из *Условий 2)–4) Определения 1* значение $D_c = 2$.

При выполнении *Условия 2)* либо *3)* число $\text{path}_C = 4$, а в множество B^C войдут признаки $m^H \in \left(M^H \backslash M^U\right)$, для каждого из которых справедливо либо соотношение (3) (при выполнении *Условия 2)*), либо соотношение (4) (при выполнении *Условия 3)*). При этом

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

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

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

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

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

где $\left(M_1^{X_1} \cup M_2^{X_1} \cup M_3^{X_1}\right) \subset M^{X_1}$ в соответствии с показанным выше разделением множества признаков формального контекста вида (1). Множество B^C здесь есть пересечение B_1 и B_2.

Значения $|B_1|$ и $|B_2|$ в формуле (6) будут тем больше, чем большее число слов могут быть синтаксически главными по отношению к каждому из слов для пары $\left(g_k^X, g^E\right)$. При этом величина $\left|B^C\right|$ отражает взаимную специфичность понятий, обозначаемых g_k^X и g^E.

Пример интерпретации теста

Пусть СЯУ S_1 задана четырьмя простыми распространенными предложениями, представляющими правильный ответ на вопрос о связи переобучения и эмпирического риска. Допустим, имеются

три варианта СЯУ S_2 (см. таблицу 1), связанные отношением схожести с S_1 согласно *Определению 1*.

Таблица 1. Сравнение ответов с эталоном.

ответы	эталон				анализируемый		
вариант	1	2	3	4	1	2	3
основа	флективная часть + предлог						
заниженн	ости	ости	ость	ость	ость	ость	ости
эмпирическ	ого	ого	ого	ого	—	—	—
риск	а	а	а	а	—	—	—
средн	—	—	—	—	ей	ей	ей
ошибк	—	—	—	—	и:на	и:на	и:на
обучающ	—	—	—	—	ей	ей	ей
выборк	—	—	—	—	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

Фрагмент тезауруса, задействованный в доказательстве схожести СЯУ, представлен в таблице 2 ЕЯ-описанием соответствующих фактов.

Таблица 2. Факты ПО для 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

Используемые в работе ФК строились по результатам синтаксического разбора ЕЯ-фраз программой «Cognitive Dwarf» (ООО «Когнитивные технологии», `http://cs.isa.ru:10000/dwarf`).

Таблица 3. Оценка близости ответа эталону.

Вариант	$\mathrm{spc}(S_1, S_2)$	$\lvert B^C\rvert$	$\lvert B_1 \backslash B^C\rvert$	$\lvert B_2 \backslash B^C\rvert$
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

Как видно из таблицы 3, наибольшее значение схожести с СЯУ S_1 имеет *вариант* 1 анализируемого ответа из представленных в таблице 1. Действительно, для этого варианта мы имеем наибольшее среднее значение $\lvert B^C\rvert$ в формуле (6) при минимальном значении суммы $\lvert B_1 \backslash B^C\rvert$ и $\lvert B_2 \backslash B^C\rvert$ по всем парам $\left(g_k^X, g^E\right)$, для которых выполняется *Условие 2), 3), либо 4) Определения 1*. Причина состоит в том, что признаки объектов формального контекста, соответствующего *варианту* 1 ЕЯ-описания ситуации S_2, разделяются большим количеством объектов формального контекста ситуации S_1, чем признаки у объектов формальных контекстов для *вариантов* 2 и 3. Иными словами, признаки для *варианта* 1 являются более стереотипическими по отношению к ФК ситуации S_1, чем признаки у двух других вариантов.

Заключение

Основной *результат* настоящей работы — *метод анализа схожести ситуаций языкового употребления при их независимом порождении.*

Предложенная теоретико-решеточная *модель тезауруса* может служить основой построения текстовых баз данных по заданной ПО. При этом классам ФП решетки тезауруса соответствуют классы СЭ в ЕЯ, за счет чего обеспечивается оптимальное иерархическое представление информации.

Немаловажную роль при вычислении меры схожести СЯУ играет полнота и непротиворечивость ЕЯ-описания предметных знаний. Модель тезауруса в виде решетки ФП позволяет задействовать, в частности, базис импликаций [1] формального контекста (2) для изучения взаимозаменяемости абстрактных слов в синтаксических контекстах существительных предметной лексики (*«связана с переобучением»*⇔*«переобучение приводит (к)»*).

Отдельного рассмотрения заслуживает интеграция предложенного метода с лингвистическими и статистическими методами поиска, используемыми алгоритмом Exactus, `http://www.exactus.ru/`.

Литература

[1] *Ganter B., Wille B.* Formal Concept Analysis — Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[2] *Михайлов Д. В., Емельянов Г. М.* Морфология и синтаксис в задаче семантической кластеризации // Всеросс. конф. ММРО-14, М.: Макс Пресс, 2009. — С. 563–566.

[3] *Останин К. С.* Система компьютерного тестирования «ТестЭкзаменатор» // Межд. конгресс конференций ИТО-2003. — 2003. — `http://www.bitpro.ru/ito/2003/VI/VI-0-2562.html`.

[4] *Челышкова М. Б.* Теория и практика конструирования педагогических тестов. Учебное пособие. — М.: Логос, 2002. — 431 с.