Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний*

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

В данной статье показывается, как можно применять методы анализа формальных понятий для оптимальной организации тестов открытой формы в системах контроля знаний. Рассматривается минимизация базы знаний, описываемых на естественном языке разработчиком теста, путем выделения смысловых эталонов.

Тестовое задание открытой формы [5] в системе контроля знаний предполагает ответ обучаемого в виде одного или нескольких предложений Естественного Языка (ЕЯ).

Как было показано нами в [4], оценка близости ответа обучаемого заданному «правильному» ответу предполагает привлечение тезауруса, формируемого на основе множеств вариантов правильных ответов по совокупности тестов заданной тематики. При этом *актуальна задача* отбора самих форм ЕЯ-описаний каждого отдельного факта предметной области для представления в тезаурусе. В настоящей работе рассматривается решение указанной задачи расширением введенной нами ранее модели Ситуации Языкового Употребления (СЯУ).

Интерпретация ответа обучаемого

Разработчик теста описывает отдельный факт некоторой предметной области множеством Семантически Эквивалентных (СЭ) ЕЯ-фраз, которые определяют СЯУ. В [4] нами использовалась модель СЯУ в виде Формального Контекста (ФК, [2]):

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

рассматриваемого в качестве информационной единицы тезауруса. Здесь 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).

Рассматривая в [4] совокупность СЯУ для известных фактов заданной предметной области как основу формирования тезауруса, авторами не накладывались какие-либо ограничения на исходное множество ЕЯ-фраз. Тем не менее, при использовании модели вида (1) в качестве единицы тезауруса ЕЯ-фразы, составляющие её основу, должны максимально точно описывать ситуацию (выражать смысл «на одном дыхании»). Ставится *задача* разделения знаний о сходных языковых формах описания различных ситуаций действительности (с одной стороны) и о внешне различающихся формах наиболее «компактного» описания каждой из ситуаций в тезаурусе (с другой стороны).

Для решения данной задачи рассмотрим единицу знаний, представляемую моделью (1) и сформированную на основе ЕЯ-фраз, отвечающих вышеуказанному требованию, в качестве смыслового эталона СЯУ. При этом введенная ранее модель СЯУ трансформируется к виду:

$$S = (T, K), \tag{2}$$

где K есть ФК вида (1) для эталона, а множество T состоит из последовательностей пар (b_i, f_i), в которых b_i соответствует основе отдельного слова в составе ЕЯ-фразы, f_i — флексии этого слова. Введем обозначения для используемых далее символьных констант: p_{fl} — для «флексия:», p_{bs} — для «главное-основа:», p_{bf} — для «главное-флексия:», для операции конкатенации — символ $\odot$, а для конкатенации через двоеточие — символ $\bullet$.

Множество T в составе структуры (2) представляет возможные формы языкового описания заданного факта действительности. В их число входят как ЕЯ-фразы, определяющие эталон СЯУ, так и не являющиеся таковыми. Для связи последних с эталоном поставим в соответствие некоторую переменную x_i каждой основе b_i, для которой существует либо признак $m \in M\colon\ m = p_{bs} \odot b_i$, либо объект $g \in G\colon\ g = b_i$. При этом на базе модели (2) строится шаблон СЯУ (верхний индекс P от англ. Pattern):

$$S^P = (d^P, T^P, K^P), \tag{3}$$

в котором все обозначения основ в составе имен объектов и признаков ФК эталона конкретной СЯУ заменяются переменными и отдельно задается список конкретизирующих четверок вида

$$\left(d^P, d^S, x_i, b_i\right), \tag{4}$$

Работа выполнена при финансовой поддержке РФФИ, проект №10-01-00146.

где d^S — идентификационный номер СЯУ, d^P — номер её шаблона.

Заметим, что в значительном числе случаев тестирования интерпретация ответа обучаемого состоит в попытке применить шаблон (3) «правильного» ответа, сформулированного разработчиком теста. При этом не требуется производить разбор ЕЯ-ответа обучаемого с привлечением внешних программ синтаксического анализа, а сама интерпретация происходит за линейное время, пропорциональное $\left|T^P\right|$.

Формирование смыслового эталона

Компоненты K^P в составе шаблонов (3) могут быть использованы для синтаксического разбора ЕЯ-фраз из ответа обучаемого. В ходе разбора строится формальный контекст вида (1) относительно некоторой фиксированной и смежных с ней предметных областей. Наличие структур-конкретизаций (4) по каждой анализируемой ЕЯ-фразе при этом не является обязательным.

Рассмотрим теперь задачу построения формального контекста самого смыслового эталона как основы моделей (2) и (3) по совокупности ФК отдельных СЭ-фраз, задающих СЯУ. Положим, что ФК указанной совокупности, далее упоминаемой как список K^{SE}, строятся по результатам синтаксического анализа этих фраз программой «Cognitive Dwarf» (ООО «Когнитивные технологии», [1]), которая использовалась нами в [4].

Для решения поставленной задачи введем коэффициенты сжатия информации относительно формальных контекстов вида (1).

Коэффициент сжатия информации по основам для формального контекста указанного вида равен:

$$k^S = \frac{\sum_{i=1}^{n^{BS}} k_i^S}{n^{BS}}, \tag{5}$$

где

$$k_i^S = \frac{\sum_{j=1}^{n_i^{BS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AS}}{n_i^{BS}},\ n^{BS} = \left|M_1\right|,\ n^{MF} = \left|M_2\right|,$$

$$n_i^{BS} = \left|\left\{\ g \in G\colon I(g,m) = \text{true},\ m \in M_1,\ m = p_{bs} \odot b_i\ \right\}\right|,$$

$$n_{ijk}^{AS} = \left|\left\{\ m_k \in M_3\colon I(g,m_k) = \text{true},\ \exists\ m_{bf} \in M_2\colon m_{bf} = p_{bf} \odot f_k,\ m_k = b_i \bullet f_k\ \right\}\right|.$$

Аналогично определяется коэффициент сжатия информации по флексиям.

$$k^F = \frac{\sum_{i=1}^{n^{FS}} k_i^F}{n^{FS}}, \tag{6}$$

где

$$k_i^F = \frac{\sum_{j=1}^{n_i^{FS}} \sum_{k=1}^{n^{MF}} n_{ijk}^{AF}}{n_i^{FS}},\ n^{FS} = \left|M_5\right|,$$

$$n_i^{FS} = \left|\left\{\ g \in G\colon I(g,m) = \text{true},\ m \in M_5,\ m = p_{fl} \odot f_i\ \right\}\right|,$$

$$n_{ijk}^{AF} = \left|\left\{\ m \in M_4\colon I(g_j,m) = \text{true},\ \exists\ m_{bf} \in M_2\colon m_{bf} = p_{bf} \odot f_k,\ m = f_i \bullet f_k\ \right\}\right|.$$

Пусть смысловые эталоны для предметно-языковых знаний эксперта фиксируются в тезаурусе, представляемом формальным контекстом:

$$K^H = (G^H, M^H, I^H), \tag{7}$$

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 указания на основы слов, синтаксически подчиненных другим словам в ЕЯ-описаниях ситуаций $g^H \in G^H$;
- множество связей «основа–флексия» для синтаксически зависимого слова;
- множество сочетаний основ зависимого и главного слова.

Отношение $I^H \subseteq G^H \times M^H$, как и $I \subseteq G \times M$ для формального контекста (1), ставит в соответствие объектам их признаки.

Положим список K^{SE} отсортированным в порядке убывания мощностей множеств объектов для входящих в него ФК. Тогда построение ФК K^E вида (1) для смыслового эталона задаётся двумя нижеприведенными алгоритмами.

Алгоритм 1. Выделение потенциальных эталонов.

Вход: K^{SE};
Выход: $P^E = \left\{K^{PE}\colon K^{PE}$ — ФК вида (1)$\right\}$;

1: взять очередной $K = (G, M, I)$ из K^{SE};
2: $N_{max}^G := |G|$;
3: $P^E := \varnothing$;
4: **для всех** $K \in K^{SE}$ таких, что $|G| = N_{max}^G$
5: $\quad K_{cur}^{SE} := K^{SE} \setminus K$;
6: $\quad K^{PE} := K$;
7: $\quad k_{max}^S := k^S\left(K^{PE}\right)$ согласно формуле (5);
8: $\quad k_{max}^F := k^F\left(K^{PE}\right)$ согласно формуле (6);
9: $\quad$ **цикл**
10: $\quad\quad$ взять очередной $K = (G, M, I)$ из K_{cur}^{SE};
11: $\quad\quad K_{cur}^{PE} := K^{PE} \cup K$;
12: $\quad\quad k_{cur}^S := k^S\left(K_{cur}^{PE}\right)$;
13: $\quad\quad k_{cur}^F := k^F\left(K_{cur}^{PE}\right)$;
14: $\quad\quad Flag := \left(\left(k_{cur}^S > k_{max}^S\right) \wedge \left(k_{cur}^F > k_{max}^F\right)\right)$;
15: $\quad\quad$ при $Flag =$ false **выход**;
16: $\quad\quad k_{max}^S := k_{cur}^S$;
17: $\quad\quad k_{max}^F := k_{cur}^F$;
18: $\quad\quad K^{PE} := K_{cur}^{PE}$;
19: $\quad P^E := P^E \cup \left\{K^{PE}\right\}$;

Замечание 1. Применительно к паре произвольных формальных контекстов $K^X = \left(G^X, M^X, I^X\right)$

и $K^Y = \left(G^Y, M^Y, I^Y\right)$ теоретико-множественная операция $K^X \cup K^Y$ понимается как построение ФК $K^U = \left(G^X \cup G^Y, M^X \cup M^Y, I^X \cup I^Y\right)$.

Для описания следующего алгоритма необходимо ввести ряд дополнительных обозначений и соглашений. Пусть *CheckAndDel* есть функция удаления из состава множества объектов каждого формального контекста в списке P^E тех объектов, которые встречаются не во всех ФК данного списка. Те признаки, которые при этом становятся не принадлежащими ни одному объекту, удаляются из множества признаков отдельного ФК функцией, обозначаемой далее Pck.

Признак будет включен в множество признаков формального контекста эталона, если он входит в пятёрку признаков $\{m_1, m_2, m_3, m_4, m_5\}$, в которой $m_1 = p_{bs} \odot b$, $m_2 = p_{bf} \odot f_1$, $m_3 = b \bullet f_1$, $m_4 = p_{fl} \odot f_2$, $m_5 = f_2 \bullet f_1$, а b есть основа некоторого слова. При этом основе b не должен соответствовать объект ФК, если есть другой объект этого же ФК, который обладает одновременно признаком m_1 и некоторым другим признаком $m = p_{bs} \odot b_1$, где $b_1 \neq b$, а основе b_1 не соответствует ни одного объекта этого ФК при том, что признак m относится более чем к одному объекту.

Функции, которая удаляет из признакового набора каждого объекта формального контекста признаки, не отвечающие данному условию, дадим имя *Closure*. Содержательно данная функция удаляет признаки главных слов-причастий в составе оборотов. Кроме того, указанная функция проверяет принадлежность каждого признака формируемого ФК $K^E = \left(G^E, M^E, I^E\right)$ множеству признаков, которые задают последовательности соподчиненных слов по следующему принципу:

$$\begin{cases} \exists\, m_1 \in M_1^E\colon \left((m_1 = p_{bs} \odot b) \wedge I^E\left(g, m_1\right)\right) = \text{true} \\ \exists\, m_2 \in M_2^E\colon \left((m_2 = p_{bf} \odot f) \wedge I^E\left(g, m_2\right)\right) = \text{true} \\ \exists\, m_3 \in M_3^E\colon \left((m_3 = b \bullet f) \wedge I^E\left(g, m_3\right)\right) = \text{true} \\ \exists\, m_5 \in M_5^E\colon \left((m_5 = p_{fl} \odot f) \wedge I^E\left(b, m_5\right)\right) = \text{true} \end{cases}$$

при максимально возможной длине каждой из последовательностей.

Замечание 2. Последовательности из трех и более соподчиненных слов, встречающиеся в 50 и более процентах СЭ-фраз из определяющих заданную СЯУ выделяются предварительно на этапе синтаксического разбора и не представлены объектами и признаками формальных контекстов из списка K^{SE} на входе *Алгоритма 1*. Для каждой такой последовательности строится свой формальный контекст вида (1), который будет объединен с формальным контекстом эталона (множество таких ФК обозначим далее как P^{SQ}). Данный шаг предпринят в целях предотвращения нежелательного занижения коэффициентов (5) и (6) при выполнении указанного алгоритма.

Будем использовать символ *Null* для обозначения формального контекста с пустыми множествами объектов и признаков. Тогда окончательный алгоритм формирования смыслового эталона будет выглядеть следующим образом.

Алгоритм 2. Формирование смыслового эталона.

Вход: P^E, P^{SQ};
Выход: K^E — ФК вида (1) для эталона;
1: $P_1^E := CheckAndDel\left(P^E\right)$;
2: $P_2^E := \left\{K_2^{PE}\colon K_2^{PE} = Pck\left(K_1^{PE}\right), K_1^{PE} \in P_1^E\right\}$;
3: $K_{tmp}^E := Null$;
4: **пока** $P_2^E \neq \varnothing$
5: взять очередной K_2^{PE} из P_2^E;
6: $K_{tmp}^E := K_{tmp}^E \cup K_2^{PE}$;
7: $P_2^E := P_2^E \setminus \left\{K_2^{PE}\right\}$;
8: $K^E := Closure\left(K_{tmp}^E\right)$;
9: **пока** $P^{SQ} \neq \varnothing$
10: взять очередной K^{SQ} из P^{SQ};
11: $K^E := K^E \cup K^{SQ}$;
12: $P^{SQ} := P^{SQ} \setminus \left\{K^{SQ}\right\}$;

Формируемый *Алгоритмом 2* смысловой эталон соответствует подмножеству максимально проективных ЕЯ-фраз исходного СЭ-множества, представляющих лучшие способы описания заданного факта действительности. Напомним, что ЕЯ-фразу следует считать проективной в содержательном смысле, если все стрелки выявленных синтаксических связей могут быть проведены без пересечений по одну сторону прямой, на которой записана эта фраза. Кроме того, если из позиции некоторого слова выходят несколько стрелок, то эту позицию не должны накрывать стрелки, выходящие из позиций других слов. Говоря о максимальной проективности, здесь мы подразумеваем минимальную суммарную длину синтаксических связей внутри ЕЯ-фразы, не превышающую длины её самой, [3].

Экспериментальная апробация

Предложенные методы формирования смыслового эталона и интерпретации ответа обучаемого были апробированы на материале ЕЯ-описаний фактов предметной области «Математические методы обучения по прецедентам». Часть указанного материала использовалась нами в [3] и [4]. При этом число СЭ-фраз, задающих СЯУ, выбиралось экспериментально с целью максиального приближения к реальной ситуации разработки теста. Данный показатель представлен в таблице 1 параметром N_1, его значение варьировалось в пределах от 2 до 54 в зависимости от описываемого факта. Для сравнения в этой же таблице приведены значения чис-

Таблица 1. Смысловые эталоны.

i	1	2	3	4	5	6
$N_1(i)$	54	53	26	26	2	3
$N_2(i)$	14	15	5	11	2	3
$N_3(i)$	13	15	13	12	8	11
$N_4(i)$	160	153	135	102	46	68
$N_5(i)$	9	12	12	12	8	11
$N_6(i)$	75	78	65	71	46	68

ла фраз, представляющих эталон (N_2), исходного числа объектов (N_3) и признаков СЯУ (N_4), числа объектов (N_5) и признаков эталона (N_6). Индекс i здесь есть порядковый номер СЯУ, краткое описание самих СЯУ даётся таблицей 2.

Таблица 2. Ситуации языкового употребления.

i	Что описывает СЯУ
1	Связь переобучения с эмпирическим риском
2	Связь переусложнения модели с заниженностью средней ошибки на тренировочной выборке
3	Влияние переподгонки на частоту ошибок дерева принятия решений
4	Причина заниженности оценки обобщающей способности алгоритма
5	Зависимость оценки ошибки распознавания от выбора решающего правила
6	Зависимость обобщающей способности логического алгоритма классификации от числа закономерностей алгоритмической композиции

Качественной характеристикой процесса формирования смысловых эталонов в целом может послужить показанное на рис. 1 соотношение размеров тезауруса (7) при формировании его на основе формальных контекстов вида (1) всех СЭ-фраз каждой СЯУ (V_1) и на основе смысловых эталонов с применением предложенных в работе алгоритмов (V_2) при заданном числе СЯУ (N), $N = \left|G^H\right|$.

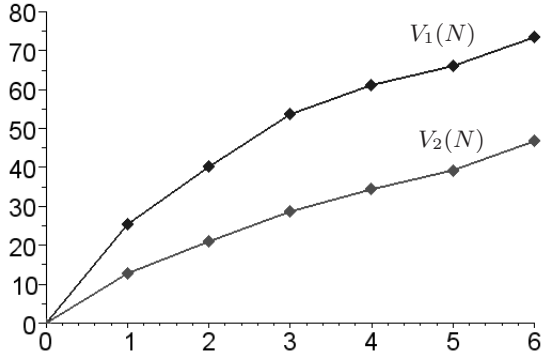

Рис. 1. Размер тезауруса для разного числа СЯУ, Кб.

Модель (7) позволяет при вычислении функции *Closure* в составе *Алгоритма 2* дополненять формируемый эталон информацией слов-синонимов по сходству лексической и флективной сочетаемости на основе ранее сформированных эталонов, представленных в тезаурусе. Именно так были выделены синонимы *«переобучение»–«переподгонка»* для СЯУ с номерами 1, 2 и 4 в таблице 2.

Заключение

Основной *результат* настоящей работы — *метод минимизации базы знаний* для вычисления рассмотренной в [4] количественной оценки схожести СЯУ при их независимом порождении. Применение предложенного метода позволяет уменьшить размер используемой базы в среднем на 40–50%.

Особенностью представленного метода является построение модели смыслового эталона по результатам разбора исходных СЭ-фраз внешней программой синтаксического анализа. Значимыми моментами здесь являются высокая точность разбора (менее 2% ошибок) для случаев существенных смысловых ограничений на перифразирование, а также свободная распространяемость таких программ (включая исходные коды), что немаловажно при построении системы тестирования знаний.

Точность описанного метода может быть оценена средним числом невыделенных (опущенных) признаков на один объект формального контекста сформированного эталона. При этом за основу оценки может быть взят аналогичный ФК, но построенный с привлечением модели процесса выявления закономерностей сосуществования словоформ в линейном ряду, предложенной нами в [3].

Тема отдельного обсуждения — формирование единого смыслового эталона для нескольких СЯУ. Практически это означает доказать возможность их включения в один класс с формированием прецедента в виде модели (1) по наличию одного из случаев синонимии, рассмотренных в [4] и составляющих основу схожести между СЯУ. Само доказательство здесь ведется относительно эталонов отдельных СЯУ, предварительно сформированных описанными в настоящей работе методами.

Литература

[1] `http://cs.isa.ru:10000/dwarf` — 2011.

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

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

[4] *Михайлов Д. В., Емельянов Г. М.* Семантическая схожесть текстов в задаче автоматизированного контроля знаний // Межд. конф. ИОИ-2010, М.: Макс Пресс, 2010. — С. 516–519.

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