УДК 658.012.56

Л.Н. Бадёрина

Метод лингвистической обработки текстових ответов в системах тестирования знаний

Рассмотрен метод оценки знаний обучающихся с помощью сравнительного анализа текстового ответа с заданным эталонным текстом и определение их релевантности, приведены практические расчеты. Исследована лингвистическая модель представления знаний на основе синонимии терминов предметной области.

The method is considered of estimating the students knowledge with the help of a comparison analysis of a text answer with the given standard text and the identification of their correspondence. Practical settlements are given. A linguistic model of the knowledge presentation on the basis of the synonymy of the terms of a subject domain is investigated.

Розглянуто метод оцінювання знань учнів шляхом порівняльного аналізу тексту відповіді з заданим еталонним текстом та визначення їх релевантності, наведено практичні розрахунки. Досліджено лінгвістичну модель уявлення знань на основі синонімії термінів предметної області.

Введение. В связи со стремительным развитием систем автоматизированного обучения актуализировалась проблема построения формальных моделей, описывающих те или иные аспекты этих систем. Среди них особенно выделяются модели и средства, ориентированные на проведение автоматизированного оценивания результатов учебного процесса. Следует отметить, что если построение учебных контентов и целостных систем в этом направлении разработаны достаточно полно, то автоматизация процессов оценивания пока находится на начальном этапе. Это связано прежде всего с тем обстоятельством, что результаты учебного процесса представляются как ответы на экзаменационные и прочие вопросы и поэтому имеют форму естественного текста. Итак, технология оценивания таким способом приобретает характер автоматического (автоматизированного) сравнения естественных языковых текстов или их фрагментов. Очевидно, что такая технология априори должна быть языкозависимой и строиться для каждого языка в отдельности. Тем более неизвестны даже общесистемные научные работы, посвященные этому предмету, что и обусловило необходимость постановки этой работы.

Общая структура системы оценивания ответов и средства моделирования

Учитывая естественноязыковую специфику предмета нашего исследования, основной теоретической конструкцией для построения модели предметной области избираем модель лексикографической среды (или интегрированной лексикографической системы), исследованной в ряде работ.

При построении модели необходимо сконструировать формальные корреляты языковых конструкций, отражающих содержание предметной области, причем моделирование необходимо выполнять как с учетом формы, так и содержания. При этом следует учитывать, что языковая система представляет собой сложную иерархию разных уровней комплексов единиц, объектов и отношений.

Первым шагом построения модели, по мнению автора, должно быть моделирование совокупности лексических единиц, отображающих «словарь» предметной области как объект исследования, поскольку именно лексической подсистеме принадлежит центральная роль в языковой системе вообще. Этот словарь должен содержать прежде всего «класс термов», состоящих из грамматически специфицированной совокупности лексем предметной области. Адекватной моделью для этого есть модель грамматической Л-системы (Г-системы), в структуре которой выделяются такие элементы:

- класс элементарных информационных единиц $V=\{x\}$, соответствующих классу всех слов

украинского языка (в рассматриваемом случае это класс термов предметной области);

• класс начальных форм, который для изменяемых частей речи соответствует исходным (словарным формам);

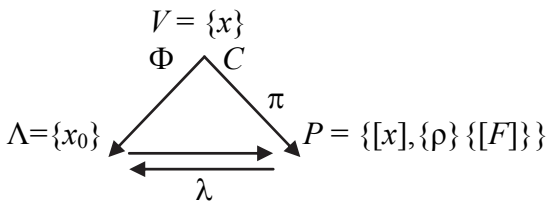
• класс разложений слов: $\pi(x) = \rho(x)*\{\omega_u(x)\}$ и соответственно множество неизменных $\{\rho\}$ и изменяющихся $[F]^k \equiv \{\omega_u(x)\}$ частей для всех слов (квазиоснов и квазифлексий соответственно);

• конечное множество словоизменяемых (парадигматических) классов: $\bigcup t_i/\pi_u$;

• оператор парадигматизации π, который ставит в соответствие каждому слову x его полную словоизменяемую парадигму $[x]$;

• оператор лемматизации λ, ставящий в соответствие каждому $\xi \in [x]$ его исходную форму x_0.

Схематично структура Г-системы представляется таким образом:



Определим на Г оператор: $\boldsymbol{R} = \pi_\rho \circ \Phi$, где π_ρ – есть ограничения π на ρ. Тогда для любого $\xi \in [x]$ справедливо:

$$\boldsymbol{R}\xi = \rho(\xi).$$

Оператор $\boldsymbol{R}$ будет использован при построении системы анализа ответов.

Итак, грамматическая система имеет такое строение:

$$\Gamma = \{V = \{x\};\ \Lambda = \{x_0\} = \Phi V;$$
$$P = CV = \pi\ \Lambda = \{[x],\ \{\rho\}\ \{[F]\};\ \lambda x = x_0;\ \boldsymbol{R}\}.$$

При этом соответственно общему определению Л-системы, справедливым будет условие: $\pi \circ \Phi = C$ и $\lambda \circ C = \Phi$, где символом «$\circ$» обозначена композиция отображений.

Функционально определенная таким способом Г-система разрешает из любого текста выделить слова (единицы лексического уровня), осуществить их грамматическую идентификацию, установить словоизменяемые классы, к которым они принадлежат, выделить для каждого слова его квазиоснову и квазифлексию, т.е. представить в форме, адаптированной для дальнейшего анализа.

Строение тезаурусной Л-системы

Построим определенную конструкцию тезауруса Σ над Г. По определению этот тезаурус строим как элементарную Л-систему, где классом элементарных информационных единиц есть класс $\boldsymbol{Z}$: $\{z\} = \boldsymbol{Z}$, $z \in \boldsymbol{Z}$ – представители реалий (понятий) рассматриваемой предметной области Σ. Формальную часть описания класса $\boldsymbol{Z}$ – совокупность определенных цепочек элементов с $V = \{x\}$ обозначим через $\boldsymbol{X}$:

$$\boldsymbol{X} = \{x_1\Delta_1\, x_2\Delta_2 \ldots \Delta_{q-1}x_q,\ q = 1, 2, \ldots\} \ldots,$$

где $x_i \in V$; символами Δ_u, $u = 1, 2, \ldots, q$, обозначен пробел (или знак пунктуации + пробел); число q – длина цепочки. Таким образом,

$$\boldsymbol{X} = \{z_q = x_1\Delta_1\, x_2\Delta_2 \ldots \Delta_{q-1}x_q,$$
$$q = 1, 2, \ldots, n, \ldots \ldots x_i \in V\} = \bigcup_{i=1} X_i$$

и $\boldsymbol{X}$ можно представить в виде объединения:

$$\boldsymbol{X} = \bigcup_{u=1} X_i,$$

где X_i – совокупность цепочек длины u. Класс цепочек длины 1 совпадает с классом $V = \{x\}$; класс цепочек длины 2 есть класс соединений из двух слов и т.д. Обозначим символом $\boldsymbol{X}^{\boldsymbol{R}}$ класс $\{\boldsymbol{R}\, z_q \equiv \boldsymbol{R}\, x_1\Delta_1\, \boldsymbol{R}\, x_2\Delta_2 \ldots \Delta_{q-1}\, \boldsymbol{R}\, x_q,\ q = 1, 2, \ldots$ $\ldots x_i \in V\} = \{\boldsymbol{R}\, z_q \equiv \rho(x_1)\Delta_1\, \rho(x_2)\Delta_2 \ldots \Delta_{q-1}\, \rho(x_q),$ $q = 1, 2, \ldots \ldots x_i \in V\}$, а символом $\boldsymbol{\mathcal{R}}$ – отображения $\boldsymbol{X}$ на $\boldsymbol{X}^{\boldsymbol{R}}$: $\boldsymbol{\mathcal{R}} : \boldsymbol{X} \to \boldsymbol{X}^{\boldsymbol{R}}$.

Тезаурус Σ над $\boldsymbol{Z}$ (обозначим $\Sigma[\boldsymbol{Z}]$) определяется таким способом.

Пусть для каждого нетривиального $z \in \boldsymbol{Z}$ (и соответствующей ему цепочки $z_q = x_1\Delta_1\, x_2\Delta_2 \ldots \Delta_{q-1}x_q$ с $\boldsymbol{X}$) определено конечное множество цепочек $C^{\Sigma}(z)$. Содержательно это множество интерпретируется как множество определений (толкований, дефиниций, интерпретаций и др.) термина z. Это обусловлено тем, что один и тот же учебный материал конкретной предметной области может быть изложен одновременно в нескольких учебниках разными авторами (экспертами), которые дают конкретным терминам свои формулировки, а потому имеет смысл говорить об экспертном наполнении тезаурусов

как своеобразных баз знаний. Каждый эксперт дает собственное формулирование того или другого термина, и таким образом, $C^{\Sigma}(z)$ представляет собой вектор:

$$C^{\Sigma}(z) = \{C^{\Sigma}{}_1(z);\ C^{\Sigma}{}_2(z);\ \ldots;\ C^{\Sigma}{}_{l(z)}(z)\},$$

где $C^{\Sigma}{}_u(z)$ – определения термина z i-м экспертом. Если для определенного z $C^{\Sigma}(z) = \varnothing$, то такие z будем исключать из $\Sigma(\boldsymbol{Z})$. С формальной точки зрения дефиниции $C^{\Sigma}{}_u(z)$, $u = 1, 2, \ldots$, также есть определенными цепочками из $\boldsymbol{X}$. Для $C^{\Sigma}(z)$ введем также эквивалентную активную формулировку, определив оператор:

$$C^{\Sigma}: z \to C^{\Sigma}(z),$$

что сопоставляет элементу тезауруса z вектор его толкований $C^{\Sigma}(z)$.

Синонимия термов и представление в лингвистической системе

При обработке естественноязыковых объектов, а тем более при их сравнении, используется ряд лингвистических фактов и отношений, с помощью которых устанавливается близость между языковыми конструкциями A и B. Общее понятие близости (сходства) формализуется довольно сложно, поэтому в прикладных системах исходят из тех определений близости, которые наиболее адекватно соответствуют содержанию задач, стоящих перед исследователями в каждом конкретном случае. В данном случае следует установить уровень усвоения студентом материала, т.е. насколько определение терминов, подаваемых студентом, *совпадают* с эталонными определениями. При этом необходимо подать количественную оценку такого совпадения, т.е. иметь соответствующую формальную модель сравнения текстовых определений терминов той или иной предметной области.

Простейшей моделью, которую можно применять в данном случае, по мнению автора, есть модель лексической синонимии. Следует исходить из того предположения, что отношение синонимии между языковыми единицами x и y, которое задается условием близости их семантических состояний $|c(x) - c(y)| < \varepsilon$, xSy, можно оценить количественно, поскольку степень синонимии между членами одного синсета может быть разной. Для удобства будем маркировать степень синонимии величиной $\delta = 1 - \varepsilon$, причем примем, что максимально возможная величина ε равняется единице, а соответственно $\delta = 0$ (когда x и y не являются синонимами), а минимальная возможная $\varepsilon = 0$, а $\delta = 1$ (когда $x = y$, или x и y есть абсолютными синонимами).

Таким образом, выполнив количественную оценку степени синонимии, в результате получим синонимическую матрицу $K(x\ y)$, $x,y \in W$, которая на формальном уровне определяется как функция с декартова произведения $W \times W$ на отрезок [0,1]. Элементы синонимической матрицы $K(x\ y)$ определяются следующим способом:

$$\left.\begin{array}{l} K(x\,x) = 1; \\ 0 < K(x\,y) \le 1;\ x \ne y;\ xSy; \\ (x\,y) = 0\ x \ne Sy. \end{array}\right\}$$

Обозначим символом $K^{\boldsymbol{R}}(x\ y)$ матрицу, которая получается из $K(x\ y)$ применением к x и y процедуры $\boldsymbol{R}$. Положим по определению:

$$K^{\boldsymbol{R}}(x\ y) = K(\boldsymbol{R}x\ \boldsymbol{R}y).$$

Это означает, что установленную экспертную оценку синонимической близости можно распространить, кроме самих термов, также и на их квазиосновы.

Релевантность понятий и их дефиниций

Распространим понятие синонимии отдельных термов, которые в лингвистическом смысле есть элементами лексической системы, на составные тезауруса предметной области $\Sigma[\boldsymbol{Z}]$. Для этого рассмотрим два аспекта – *формальный* и *содержательный*.

С *формальной точки зрения* задача состоит в установлении семантической близости, аналогичной к свойству синонимии, но не на множестве отдельных термов, а на множестве цепочек вида $x_1\Delta_1\ x_2\Delta_2 \ldots \Delta_{q-1}x_q$, $q = 1, 2, \ldots$, определенных формулой при условии, что элементы $x_1,\ x_2, \ldots, x_q$ попадают в область определения функции $K(x\ y)$. *Содержательный* аспект предусматривает установление отношения семантической близости, аналогичной свойству синонимии, на множестве дефиниций сроков:

$$C^{\Sigma}(\boldsymbol{Z}) = \{C^{\Sigma}(z) \mid {}_{\forall z \in \Sigma(z)}\} = \{\{C^{\Sigma}{}_1(z);\ C^{\Sigma}{}_2(z);\ \ldots;$$

$C^{\Sigma}_{l(z)}(z)\} \mid_{\forall z \in \Sigma(z)...}\}$. Поскольку понятия синонимии в лингвистике корректно определяются лишь для лексической системы, то для установления содержательной (семантической) близости элементов с $C^{\Sigma}(z)$ введем понятие *отношения релевантности*, которое обозначим символом ***REL***.

С этой целью определим количественную меру релевантности двух цепочек: $A = z$ и $B = z$ (длиной M и N соответственно), которую обозначим как ***REL***(A,B).

Таким образом, определяется отображение ***REL***: $C^{\Sigma}(\boldsymbol{Z}) \times C^{\Sigma}(\boldsymbol{Z}) \to \Delta$, где Δ – определенное подмножество множества неотрицательных чисел. При этом будем считать, что цепочка B релевантна цепочке A, т.е. A ***REL*** B, тогда и только тогда, когда значение функции ***REL***(A,B) не меньше какого-либо определенного $\delta \in \Delta$: ***REL***$(A,B) \geq \delta$, выбор которого зависит от специфики предметной области и конкретных задач исследования и оценивания.

Для нахождения явного вида меры релевантности используем такую аналитическую модель: ***REL***$(A,B) = \omega\eta$, где ω – определенная функция, зависящая от числовых значений показателя синонимии термов, которые входят в A и B, т.е. определенной функцией матричных элементов матрицы синонимии $K(x, y)$; η – есть определенная функция от длин цепочек A и B (т.е. от M и N).

Функция η задает зависимость меры релевантности ***REL***(A, B) от количества термов в цепочках A и B, т.е. от целых чисел M и N. Очевидно, что только при совпадении цепочек, они максимально релевантны и следовательно тогда ***REL***(A, B) достигает максимального значения. Из этого следует первое свойство η:

1. $\eta = \eta_{\max}$ тогда и только тогда, когда $M = N$.

Также очевидно, что функция η симметрична по переменным M и N, следовательно, она симметрична относительно своего максимального значения. Из этого следует второе свойство η:

2. $\eta\ (M, N) = \eta\ (N, M)$ и η симметрична относительно значения $\eta_{\max}$.

Простейшей функцией с такими свойствами есть функция от $|M - N|$.

Следующее свойство функции η связано с ее поведением при относительно больших различиях $|M - N|$. Очевидно, что если цепочки A и B сильно отличаются по длине (количестве значимых термов), то они не могут быть релевантными, поскольку каждый терм, отсутствующий в одной цепочке, вносит свой вклад в семантику второй цепочки, и с учетом каждого такого терма различие в семантике цепочек усиливается, а их взаимная релевантность – снижается. Из приведенных соображений вытекает третье свойство функции η:

3. Если $|M - N| \to \infty$ (или $M - N \to \pm\infty$), то $\eta \to 0$.

Условия 1–3 задают определенное функциональное уравнение, одним из решений которого есть функция $\eta\ (M, N)$, которая имеет вид

$$\eta(M, N) = l(h)e^{-h|M-N|^2},\ h > 0.$$

Параметр h и функцию $l(h)$ подбирают экспериментально, и они могут варьироваться по условиям пользователя; они также могут определяться путем наложения определенных условий нормирования.

Функция ω зависит от трех сменных: коэффициента синонимии термов $K\ (A, B)$ в цепочках A и B, количества значащих термов в цепочке A (т.е. от N) и количества значащих термов в цепочке B (т.е. от M): $\omega = \omega\ (K\ (A, B); N; M)$.

Количество термов в цепочке A, очевидно, равняется мощности множества $\hat{A}$. Количество термов в цепочке $\hat{A}$ равняется мощности множества $\hat{A}$ до обработки:

$$N = \|A\|, M = \|B\|.$$

Если $N = \text{const}$, то функция ω имеет такие свойства:

1. При увеличении $K\ (A, B)$ значение ω возрастает.

2. С увеличением количества термов ответа значение ω снижается.

3. Если $\frac{\sum_{e \in B} k_t}{M} = 1$, то $\omega = \omega_{\max}$. В этой формуле величины k_t – определенные элементы матрицы синонимичности для так называемых

нормализованных цепочек, определению которых будет посвящена отдельная работа.

Обобщенный параметр $K(A, B)$ равняется сумме максимальных коэффициентов соответствия термов цепочек A и B: $K(A, B) = \sum_{e \in B} k_e$.

На основании приведенных свойств получена формула для представления ω:

$$\omega = \frac{\sum_{e \in B} k_t}{M}.$$

Заключение. Мера релевантности цепочек A и B как произведение функций ω и η, фигурирующих в формуле, определяется так:

$$\boldsymbol{REL}(A, B) = \omega\eta = \frac{\sum_{e \in B} k_t}{M} . l(h) e^{-h|M-N|^2}, \quad h > 0 .$$

Эта формула на самом деле учитывает определенные эффекты семантической близости информационно-языковых объектов, поэтому ее можно применять как инструмент при анализе ситуаций, возникающих при сравнении эталонных (представленных в нормативных источниках, в частности учебниках) формулировок понятий и дефиниций предметной области с фактическими их формулировками, которые есть объектами оценивания, если и первые и вторые представлены цепочками вида A и B.

Описанная модель была применена в процессе оценивания реальных ответов студентов курса «Информатика и вычислительная техника».

1. *Kinshuk D.*, *Patel A*. A conceptual framework for Internet based intelligent tutoring systems. Knowledge transfer (**II**) // Educational Technology & Society. – P. 117–124. – http://ifets.ieee.org
2. *Кириличев Б.В.*, *Широков Л.А.*, *Рабинович П.Д.* Системный анализ проблемы создания интеллектуальных компьютерных обучающих комплексов: Сб. науч. тр. МГИУ. – Г.: МГИУ. – 1996. – С. 166–171.

Поступила 28.12.2010

Тел. для справок: (067) 442-2688 (Киев)
E-mail: *vada@ukr.net*

Окончание статьи С.Д. Погорелого и др.

При учете всех накладных временных расходов на выделение памяти видеоадаптера, копирование данных на и от него, создание миллионов потоков, синхронизацию потоков на k-й итерации центральным процессором, предложенная параллельная реализация на видеоадаптере минимум в два раза быстрее последовательной реализации на центральном процессоре.

Использование разделяемой памяти графического процессора не дает преимуществ без снижения количества обращений к глобальной памяти видеоадаптера, в которую копируется матрица весов графа.

Для дальнейшего сокращения времени выполнения алгоритма на видеоадаптере необходимо использовать иные подходы к выделению информационных зависимостей, которые, возможно, теоретически дадут больше время выполнения алгоритма, чем $O(N)$, однако, при реализации, используя программно-аппаратную платформу *CUDA*, будут использовать разделяемую память и снизят число обращений в глобальную память видеоадаптера.

1. *Кормен Т.*, *Лейзерсон Ч.*, *Ривест Р.* Алгоритмы, построение и анализ. – М.: МЦНМО, 2000. – С. 719–725.
2. *Анализ* методов повышения производительности компьютеров с использованием графических процессоров и программно-аппаратной платформы *CUDA* / С.Д. Погорелый, Ю.В. Бойко, М.И. Трибрат и др. // Математичні машини та системи. – 2010. – № 1. – С. 40–54.
3. *Методика* вимірювання обчислювальної потужності відеоадаптера (платформа *CUDA*) / С.Д. Погорілий, М.И. Трібрат, Ю.В. Бойко та ін. // ИКВТ (ДонНТУ), 2010. – № 11. – С. 94–98.
4. *Погорілий С.Д.*, *Камардіна О.О.*, *Бавикін О.І.* Про підхід до розпаралелювання алгоритму Флойда–Уоршала // Математичні машини і системи. 2005. – № 3. – С. 91–101.
5. *NVIDIA* CUDA Programming Guide 2.3. – http://developer.download.nvidia.com/compute/cuda/2_3/toolkit/docs/NVIDIA_CUDA_Programming_Guide_2.3.pdf
6. *Программное* обеспечение *UAClaster* / В.А. Мар'яновский, С.Д. Погорелый, Ю.В. Бойко и др. // УСиМ. – 2009. – № 5. – С. 76–80.
7. *Inter-Block* GPU Communication via Fast Barrier Synchronization. – http://www.nvidia.com/content/GTC/posters/ 73_Feng_Accelerating_Applications.pdf
8. *PCI Express*. –http://ru.wikipedia.org/wiki/PCI_Express
9. *All-Pairs* Shortest-Paths for Large Graphs on the GPU / G.J. Katz, J.T. Kider Jr. – http://www.seas.upenn.edu/~kiderj/ research/papers/APSP-gh08-fin-T.pdf

Поступила 07.12.2010

Тел. для справок: (044) 526-0522 (Киев)
E-mail: *sdp@univ.kiev.ua*, *mike3b@univ.kiev.ua*, *boyko@univ.kiev.ua*, *dima@univ.kiev.ua*