О ВОЗМОЖНЫХ НАПРАВЛЕНИЯХ ФОРМАЛЬНОГО АНАЛИЗА КАЧЕСТВА ТЕСТОВЫХ МАТЕРИАЛОВ ДЛЯ КОНТРОЛЯ ЗНАНИЙ

Юрий Брумштейн, Светлана Окладникова

***Аннотация**: ...*

Использование тестов является важной компонентой учебного процесса и систем профессиональной аттестации специалистов. Общий рост объемов тестовых материалов (ТМ) и расширение их номенклатуры стимулируются: автоматизацией и компьютеризацией процессов обучения в вузах; распространением системы дистанционного обучения и пр.

Программные средства автоматической генерации ТМ на основе содержания учебных курсов обычно дают неудовлетворительные результаты, т.к. механическая замена повествовательных предложений на вопросы ТЗ оказывается недостаточной. Поэтому ТМ создаются вручную, этот процесс носит творческий характер, а его результаты могут рассматриваться как объекты интеллектуальной собственности. Качественные и востребованные ТМ могут быть объектами коммерциализации. Расширение связей между странами в области высшего образования, международная унификация процессов сертификации специалистов и пр. заставляют рассматривать ТМ и как потенциальный объект международного обмена. В силу указанных причин разработка объективных методов оценки качества ТМ представляется важной и актуальной задачей.

Далее под "тестом" будем понимать совокупность тестовых заданий (ТЗ). В общем случае на основе одной и той же базы ТЗ (БТЗ) могут быть сформированы различные тесты, в том числе и разной сложности. Отдельные ТЗ и БТЗ в целом могут оцениваться: (**А**) на основе некоторых формальных показателей (критериев) до начала фактического использования ТЗ (т.е. на стадии разработки); (**Б**) методами "экспертного оценивания" компетентными специалистами (как до применения, так и по его результатам); (**В**) экспериментально - по результатам предварительного тестирования достаточно больших групп студентов и других лиц.

Преимущества методов группы "А": применимость на стадиях разработки ТЗ и при оценке целесообразности использования; удобные возможности автоматизации, что существенно при необходимости оценки ТМ большого объема; использование этих методов для ТМ по различным дисциплинам без участия экспертов; инвариантность по отношению к языку ТЗ. Основной недостаток методов этой группы - отсутствие (или ограниченные возможности) анализа ТЗ в содержательном плане. В данной работе авторы сосредоточили внимание на методах группы "А" (методы других групп требуют отдельного рассмотрения).

Важнейшими показателями качества ТЗ в рамках анализа по формальным показателям выступают: (**П1**) сложность формулировок ТЗ; (**П2**) соответствие отдельных ТЗ и БТЗ теме(ам) курса; (**П3**) полнота охвата БТЗ совокупности тем по курсу. Эти показатели могут применяться отдельно и в совокупности. В частности показатель "П1" важен для процессов адаптивного тестирования и обучения.

Типичными в практике работы вузов являются ТЗ состоящие из: сформулированного текстового вопроса; предопределенного набора ответов (также преимущественно текстового характера), из которых надо выбрать все верные или неверные (в зависимости от формулировки вопроса ТЗ). По умолчанию ниже мы

будем иметь в виду именно такой вариант ТЗ, однако рассматриваемые подходы частично применимы и для "ТЗ открытой формы".

Укажем частные критерии для формальной оценки "сложности" отдельных ТЗ.

(**C_1**) Суммарная длина формулировки вопроса и всех ответов ТЗ, выраженная как "количество символов", включая пробелы внутри фраз и некоторые знаки препинания (см. пункт C_2) . Фактически "C_1" - мера объема текста ТЗ.

(**C_2**) Количество знаков препинания и скобок в формулировках в расчете, например, на 10 слов ТЗ (целесообразен относительный, а не абсолютный показатель). Мы имеем в виду такие знаки препинания как: запятая; двоеточие; точка с запятой; тире. Нецелесообразно учитывать: дефис внутри составных слов типа "какое-либо"; знаки переноса слов; символы, расположенные в конце фраз – точку, вопросительный и восклицательный знаки.

В качестве единых "элементов для подсчета" будем рассматривать: парные скобки (как правило, круглые); парные кавычки (они обычно свидетельствуют о применении термина в переносном смысле, что усложняет восприятие текста). Итак, "C_2" - это мера сложности конструкций фраз в ТЗ.

(**C_3**) Средняя длина предложений в ТЗ (суммарно - по формулировке вопроса и набору всех ответов) выраженная в количестве слов. Длинные фразы воспринимаются с трудом, поэтому "C_3" также может рассматриваться как показатель сложности ТЗ.

(**C_4**) Средняя длина слов во всех предложениях, составляющих ТЗ (более длинные слова усложняют восприятие). Для оценки целесообразно "не считать" все предлоги и союзы ("и", "или" и пр.). Показатель "C_4" также мера сложности ТЗ. Он может определяться: предметной областью, для которой предназначено ТЗ (и, как следствие, ее терминологической базой); стилем автора-разработчика ТМ; его словарным запасом.

(**C_5**) Количество ответов в ТЗ (чем их больше, чем ТЗ обычно сложнее). Некоторые из показателей C_1…C_5 могут коррелировать друг с другом, поэтому мы предполагаем провести оценку коэффициентов корреляции на реальных ТМ.

(**C_6**) Совпадение языка ТМ с родным языком тестируемых лиц (работа с тестами на иностранном языке практически всегда требует больших усилий).

Помимо частных критериев сложности могут быть использованы и интегральные (I_n). Их целесообразно конструировать как линейные комбинации перечисленных выше частных критериев C_1…C_5 с различными весовыми коэффициентами, а C_6 рассматривать как общий множитель.

$$I_n = (C_6) * \sum_{m=1}^{5} K_{m,n} C_m \; ; \qquad \sum_{m=1}^{5} K_m = 1 \qquad (1),$$

где ***n*** - индекс интегрального критерия; { $K_{m,n}$ } – система весовых коэффициентов для ***n***-ого интегрального показателя (часть из них может быть и нулевыми).

При разных наборах { $K_{m,n}$ } будет осуществляться разная акцентуация частных критериев сложности, поэтому и интерпретация разных интегральных критериев I_n будет различной.

Анализ в отношении "C_1…C_5" может производиться по каждому из ТЗ, а полученные результаты затем обрабатываться статистически. Например, могут определяться: средние значения по БТЗ для тестов по какому-то курсу; коэффициенты вариации (КВ) для сложностей ТЗ в этой базе и пр. Высокие значения КВ

могут говорить о внутренней неоднородности БТЗ. В этом случае ТЗ с резко отличающимися показателями из БТЗ целесообразно либо исключить, либо переработать.

Параметр C_6 объективно является свойством не самого ТЗ, а условий его применения (фактически для каждого из тестируемых, т.к. степень владения ими иностранным языком может различаться). Однако в практическом плане C_6 может быть полезен. Если язык ТЗ - родной для тестируемых, то его следует принять равным "1", в противном случае взять большим "1" (усреднено для предполагаемых к тестированию групп). Величину "C_6" можно оценить экспериментально: типичная группа разбивается на две подгруппы случайным образом; одинаковые наборы ТЗ одной из подгрупп предъявляются на родном языке, а другой – на иностранном; C_6 – оценивается с учетом соотношения качества ответов и затраченного времени для этих двух подгрупп.

База ТЗ может быть и специально разработана так, чтобы в ней имелись вопросы разной сложности (например, предназначенные для "адаптивного тестирования" с динамическим подбором ТЗ по сложности в процессе тестирования). Тогда высокие КВ для показателей сложности не должны рассматриваться как недостаток.

Если базы ТЗ специально сегментируются на группы вопросов разной сложности (в т.ч. и для целей подбора ТЗ при адаптивном тестировании), то показатели КВ целесообразно применять к отдельным группам. Разницы между группами ТЗ в отношении средних показателей сложности для них могут использоваться для оценки качества "ручной" разбивки на группы. Например, все ТЗ вручную разбиты на три группы по сложности: (1) высокая, (2) средняя и (3) низкая. Желательно, чтобы разницы средних сложностей (либо по некоторым частным критериям, либо по интегральным) между 1 и 2 группами и 2 и 3-ей группами были примерно равны.

Важным является вопрос сравнительной оценки разных БТЗ в отношении средней сложности. При этом обычно, каких-то утвержденных эталонов сложности для БТЗ - нет (по крайней мере, в вузах). Дополнительно отметим: при переходе от младших курсов к старшим (от 1-го к 5-ому) максимально допустимая сложность БТЗ может увеличиваться; предельная сложность БТЗ по одному и тому же курсу для разных специальностей/факультетов (и разных форм обучения) также может серьезно различаться. Поэтому в качестве ориентира по сложности (как в отношении частных критериев, так и интегральных) может быть выбрана средневзвешенная сложность баз ТЗ (D^*) для определенной специальности и формы обучения.

$$D^* = (\sum_{j=1}^{J} N_j D_j) / (\sum_{j=1}^{J} N_j) \qquad (2),$$

где D_j и N_j - соответственно оценка сложности и количество ТЗ для j-ой БТЗ.

Для автоматизированной группировки ТЗ по сложности могут быть продуктивны методы многомерного кластерного анализа (по совокупности частных показателей сложности). Однако с учетом их корреляции, будет корректным проводить кластеризацию непосредственно по показателям $C_1 \ldots C_5$. В качестве альтернативного варианта можно использовать метод главных факторов и кластеризацию ТЗ по 1-му и 2-му главным факторам.

Выделенные кластеры ТЗ по сложности целесообразно сопоставлять с темами курса, для которого предназначается БТЗ (поскольку сами по себе темы могут обуславливать усложнение формулировок, в том числе и в силу терминологической базы).

Кластерный анализ может быть применен и для сопоставления БТЗ, предназначенных для различных учебных курсов. При этом могут использоваться два варианта: отображение БТЗ в системе двух

ортогональных осей (например по параметрам "C_1" и "C_2") кружками одного размера; аналогично, но диаметры кружков - пропорциональны третьему параметру).

Оценка соответствия ТЗ изучаемому курсу должна производиться на основе семантического анализа текстов ТЗ и их сопоставления с формулировкой названия курса.

Однако название курса обычно краткое. Поэтому, как минимум для каждого из слов в названии дополнительно должны быть использованы еще наборы "слов-синонимов". Применение для этой цели тезауруса Microsoft Word (и аналогичных текстовых процессоров) чаще всего не будет продуктивным из-за отсутствия лексики для узких предметных областей. Обычно даже добавление синонимов будет являться недостаточным для получения полноценного терминологического профиля.

Укажем, что для всех слов должна быть учтена возможность их представления в различных словоформах (и в ТЗ, и в названии курса). Разнообразие словоформ определяется следующими характеристиками: "единичный - множественный"; залог; "женский - мужской род"; падеж (влияет на падежное окончание).

Перспективными объектами, на основе которых может быть непосредственно построен "терминологический профиль учебного курса" (ТПУК), можно считать следующие тексты: (**а**) перечень формулировок основных тем по курсу, обычно приводимый в рабочей программе; (**б**) тот же перечень, дополненный перечнем пунктов для каждой из тем; (**в**) пояснительная записка-обоснование к курсу; (**г**) набор вопросов к экзамену или зачету по данному учебному курсу; (**д**) стандартный учебник соответствующей предметной области, представленный в электронной форме и допускающий перевод в формат *.txt; (**е**) – оглавление стандартного учебника, если оно достаточно подробное (его можно отсканировать с бумажного варианта учебника, а затем распознать как текст).

Терминологический профиль ТЗ (ТПТЗ) строится на основе совокупности слов в ТЗ (тексты вопроса и всех ответов). Целесообразность при построении ТПУК и ТПТЗ исключения из текстов предлогов и союзов нуждается в специальном обсуждении.

Полученные наборы терминов для ТПУК и ТПТЗ, как правило, будут различаться – и по наборам терминов и по их количеству (предполагаем, что термины, встречающиеся в разных словоформах, приведены к единым "эталонам" для них). Перейдем к "расширенным наборам" (в них включены все термины, встречающиеся в ТПУК и/или в ТПТЗ). Длины рядов терминов в "расширенных наборах" для учебного курса и ТЗ будут, очевидно, совпадать.

Повторяемость терминов в ТЗ и объекте, по которому строится ТПУК, учтем, рассчитав "долю их встречаемости" в соответствующих текстах. Вектора долей встречаемости и будем считать характеризующими ТПУК и ТПТЗ. Степени их соответствия целесообразно оценивать через парные коэффициенты корреляции - обычный (по Пирсону) и ранговый (по Спирмену). С нашей точки зрения первый из них предпочтителен.

Поскольку отдельные ТЗ часто ориентированы лишь на достаточно узкие вопросы, то показатели совпадения ТПТЗ для отдельных ТЗ и ТПУК могут быть невысокими. Поэтому, если ТЗ является узкотематическим, то более обоснованно сопоставление ТПТЗ с терминологическим профилем такого объекта учебного курса, как "название темы + названия всех пунктов темы". Автоматически определять к какой именно теме относится такое ТЗ проще всего, оценивая совпадения с каждой из тем, и выбирая в качестве "профильной" ту, для которой совпадение максимально и превышает некоторое пороговое значение (если такого превышения нет, то ТЗ следует считать "политематическим"). Альтернативный подход – в качестве "составного объекта" для построения ТПУК берутся все темы, для которых степени совпадения с ТПТЗ превышают пороговые значения. В качестве последних можно брать, например, половину среднего арифметического значения совпадений со всеми темами.

Больший интерес (по сравнению с отдельными ТЗ) представляет сравнение с ТПУК терминологического профиля БТЗ (ТПБТЗ) по курсу. Причины: значительно больший объем текста и предполагаемый всесторонний охват курса. ТПБТЗ может быть построен по тому же алгоритму, что и ТПТЗ. В этом случае степень совпадения терминологических профилей фактически оценивает не только соответствие БТЗ курсу, но и частично полноту тематического охвата. Специально для оценки последнего параметра могут быть сформулированы и некоторые другие критерии (на основе векторов совпадений отдельных ТЗ с отдельными темами курса – см. выше).

Эталонные значения для совпадений ТПУК и ТПБТЗ обычно отсутствуют. Поэтому при наличии в вузе достаточного количества БТЗ можно оценить некоторое среднее значение для совпадений и дополнительно обратить внимание на БТЗ, у которых совпадение существенно ниже среднего.

Сравнение двух различных БТЗ на совпадение терминологических профилей может иметь смысл при одновременном сравнении профилей для отвечающих им объектов, по учебным курсам (например, стандартных учебников). Оптимальным вариантом будет, очевидно, являться такой, когда в обоих случаях степени совпадений будут одинаковыми или близкими.

Итак, авторами статьи рассмотрены некоторые направления формального анализа текстов тестовых заданий в отношении характеристик их качества. Предложен ряд алгоритмов, позволяющих производить такой анализ в автоматизированном режиме. В дальнейшем предполагается опробовать эти алгоритмы на реальных тестовых материалах.

Литература

[Нейман Ю.М, 2000] Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 169 с.

[Никифорова А.М., 2000] Никифорова А.М. Способы оценки тестовых заданий в системе дистанционного обучения «KnowledgeCT» //Материалы Третьей Всероссийской научной конференции молодых ученых и аспирантов. Новые информационные технологии. Разработка и аспекты применения. Тезисы докладов. Таганрог: ТРТУ, 2000, с. 167-168.

[Сысоева Л.А., 2003] Сысоева Л.А. Методика построения логико-семантической модели структуры содержания дисциплины/Вопросы тестирования в образовании. № 8, 2003. – Москва, Центр тестирования при Министерстве образования РФ. 2003. С. 15-20 с.

[Челышкова М.Б., 1996] Челышкова М.Б. Разработка педагогических тестов на основе современных математических моделей. /Уч. пособие. – М.: Исслед. центр проблем качества подготовки специалистов, 1995. 32 с.: ил.

Информация об авторах

Брумштейн Юрий Моисеевич, к.т.н., доцент кафедры «Управление качеством» АГУ; e-mail: *brum2003@mail.ru*

Окладникова Светлана Владимировна, ведущий программист Регионального ресурсного центра дистанционного обучения АГУ; e-mail: chelle@aspu.ru

Астраханский государственный университет; 414056, Россия, г. Астрахань, ул. Татищева, 20а; Тел. (8512)61-08-11, факс (8512)54-90-99; *www.aspu.ru*