

8. Website. Wikipedia, the free encyclopedia. – [website]. – <http://en.wikipedia.org/wiki/Website>
9. Пелешишин А.М. Позиціонування сайтів у глобальному інформаційному середовищі (монографія) / Львів: Видавництво Львівської політехніки, 2007. 260 с.
10. Lampson B. Authentication in distributed systems: theory and practice // Lampson B., Abadi M., Burrows M., Wobber E. / ACM Transactions on Computer Systems (TOCS). – Nov. 1992. – v.10 n.4. – P.265-310.
11. Fernandez E.B. An authorization model for a shared database // Fernandez E.B., Summers R.C., Coleman C.D. / Proceedings of the ACM-SIGMOD International Conference, 1975 – ACM, New York.
12. Rabitti F. A model of authorization for next-generation database systems // Bertino E., Kim W., Woelk D. / ACM Transactions on Database Systems (TODS). – March 1991. – v.16 n.1. – C.88-131.
13. Gibson R.K. Online Participation in the UK: Testing a 'Contextualised' Model of Internet Effects // Gibson R.K., Lusoli W., Ward S. / The British Journal of Politics and International relations (BJPIR) . – 2005. – vol.7. – P.561-583.
14. Recordon D. OpenID 2.0: a platform for user-centric identity management // Recordon D., Reed D. / Proceedings of the second ACM workshop on Digital identity management. – November 03-03, 2006, Alexandria, Virginia, USA. – P.11-16.
15. Disqus, a service and tool for web comments and discussions. – [website]. – <http://disqus.com/docs/about/>
16. The top 500 sites of the Web. Alexa, The Web Information Company. – [website]. – <http://alexa.com/topsites>
17. The top 100 sites in Ukraine. Alexa, The Web Information Company. – [website]. – <http://alexa.com/topsites/countries/UA>
18. Рубрика "Україна". Яндекс Каталог. – [Електронний ресурс]. – <http://yaca.yandex.ru/yaca/geo/CIS/Ukraine/>
19. Рубрика "Ukrainian". DMOZ, Open Directory Project. – [Електронний ресурс]. – <http://www.dmoz.org/World/Ukrainian/>

Розглянуто проблему тестування теоретичних знань в автоматизованих системах контролю. Запропоновано методи тестування теоретичних знань на основі використання тезауруса семантичних полів для побудови концептуальних моделей тексту відповіді на природній мові

Ключові слова: тестування знань, відповідь на природній мові, тезаурус

Рассмотрена проблема тестирования теоретических знаний в автоматизированных системах контроля. Предложены методы тестирования теоретических знаний на основе использования тезауруса семантических полей для построения концептуальных моделей текста ответа на естественном языке

Ключевые слова: тестирование знаний, ответы на естественном языке, тезаурус

The problem testing of theoretical knowledge by automated testing systems is considered. Methods of construction of semantic models of natural language answer by means of the thesaurus of semantic fields are proposed

Key words: testing of knowledge, natural language answer, thesaurus

УДК 681.3:519.76

КОНТРОЛЬ ЗНАНИЙ С ОТВЕТАМИ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Ю. Ю. Черепанова

Ассистент

Кафедра ПОЭВМ

Харьковский национальный университет

радиоэлектроники

пр. Ленина, 14, г. Харьков, Украина, 61166

Контактный тел.: 8 (057) 402-14-46

E-mail: cher_y@list.ru

1. Введение

Эффективность любой системы образования во многом определяется качеством средств контроля знаний обучаемого. Контроль необходим как при предварительном тестировании для определения начального

уровня знаний, так и после прохождения курса обучения для контроля усвоения материала. Это касается и автоматизированных обучающих систем, и самообразования, и традиционного образования, основанного на взаимодействии «учителя» и «ученика». В связи с возрастанием роли дистанционного обучения зада-

ча повышения эффективности и интеллектуальности средств анализа ответов стала особенно актуальной. Особенности дистанционного образования объясняются новыми требованиями к системам контроля.

Во-первых, диалог между системой и обучаемым должен идти на языке, максимально приближенном к естественному. Человечество выражает свои знания вербально. Каково бы ни было внутреннее представление знаний, но о них всегда можно рассказать. Сохраняются знания и передаются следующим поколениям также в виде написанных текстов. Поэтому можно говорить, что наиболее эффективным контроль знаний будет тогда, когда диалог между системой и обучаемым будет происходить на естественном языке.

Во-вторых, наряду с тестированием практических знаний, предметом такого диалога должна быть «теория» – те знания, которые выражаются на естественном языке и представляют собой основные понятия, концепции, законы той области знаний, которая изучалась и по которой идет тестирование.

В-третьих, в связи с высокой скоростью обновления знаний, система контроля должна быть способна к «обучению» - адаптации к новым знаниям, то есть должны быть усовершенствованы функции, облегчающие ее наполнение. Перечисленные требования выдвинуты на основе анализа проблем тестирования знаний при дистанционном обучении, однако, они не теряют актуальности и в других системах образования.

2. Анализ существующих методов тестирования теоретических знаний

Рассмотрим существующие методы тестирования теоретических знаний, понимая под теоретическими знаниями сведения, выраженные на естественном языке, и содержащие основные теоретические положения в какой-либо области знания.

Одним из наиболее легко реализуемых и потому очень часто применяемых является метод тестирования с использованием выборочных ответов [1-3]. Обучаемому предоставляется вопрос и альтернативы ответов, из которых необходимо выбрать один или несколько правильных. Достоинствами данного метода является то, что он обладает высокой реактивностью, позволяет экономить память и трудозатраты со стороны автора обучающего курса. Однако этот метод обладает и недостатками. Велика вероятность просто случайного выбора правильного варианта. Таким образом, тестирование сводится к угадыванию, интуитивному определению правильного ответа. Кроме того, даже человек, сознательно выбравший верный ответ, не всегда правильно сформулирует нужное теоретическое положение. А между тем, наиболее ценны и осознанны те знания, которые человек может вербализовать. И речь идет не о том, чтобы заученно воспроизвести нужный отрывок из книги, а суметь изложить основные положения теории в свободной форме, не обязательно в «книжной» формулировке.

Эта свободная формулировка является ограничением для применения другого метода анализа ответов, когда для каждого вопроса в систему закладываются все возможные шаблоны ответа на предметно-ограниченном естественном языке, и правильность получен-

ного от тестируемого ответа определяется наличием его в списке верных шаблонов. Недостатком такого подхода является то, что количество возможных конструкций даже на ограниченном естественном языке очень велико и трудно их все предугадать, поэтому велика вероятность задания неполного списка правильных шаблонов. Кроме того, даже полное перечисление всех возможных вариантов ответа неоправданно увеличивает затраты ресурсов - памяти компьютера и времени автора. Чтобы ввести в систему контроля знаний вопрос, разработчик программы (или эксперт, ее наполняющий) должен перечислить все возможные ответы на него, и таких же трудозатрат требует каждый следующий вопрос.

Не всегда дают полную картину оценки знаний и правильное выполнение заданий с ограничениями на ответ, когда диалог идет на естественном языке, но для ответа требуется ввести строго ограниченные слова. К этому виду ответов относятся такие, которые представляют собой лишь заполнение некоторых пропущенных параметров в предоставленном системой готовом тексте ответа. Другим примером использования фиксированно-конструируемых ответов являются однословные ответы, например в том случае, когда обучаемому в вопросе предъявляется формулировка определения и для ответа необходимо привести термин. Такие методы в большей степени проверяют ассоциативную память, а не знания по теме вопроса. То, что человек по определению термина или описанию явления даст правильный термин или название явления, не всегда означает обратное.

Наибольшее распространение в системах контроля знаний находит универсальный и простой метод анализа по ключевым словам [1, 2], когда в эталоне ответа на вопрос указываются слова и синонимы, которые должны присутствовать в ответе. Чаще всего при этом не учитывается порядок слов. Данный метод обладает высокой реактивностью и простотой реализации, но не позволяет распознавать смысл текста. Кроме того, возникает та же проблема перечисления всех слов, которые могут входить в правильные конструкции ответа. При чрезмерном ограничении естественного языка, недостаточном количестве вариантов возможна неадекватная оценка правильного ответа, при избыточном количестве вариантов из-за большого числа ключевых слов велика вероятность признания ошибочной конструкции языка как правильной.

3. Цель

Для более адекватной оценки знаний обучаемого возникает необходимость в нахождении новых методов тестирования знания теоретического материала с высокой реактивностью, простотой реализации и минимальными трудозатратами автора обучающего курса. Наилучшим методом представляется способ тестирования, при котором проводится анализ смыслового соответствия естественно-языкового ответа обучаемого и теории с целью установления степени его адекватности. При этом минимальными трудозатратами автора будут в том случае, если теория будет представлена в системе в виде естественно-языковых текстов

либо модели, которую возможно построить по таким текстам с наиболее полной автоматизацией. Следовательно, возникает необходимость в механизме построения смысловой модели по естественно-языковому тексту. В настоящее время еще не существует полного решения данной задачи, но даже частное решение имеет реальный практический интерес.

4. Построение моделей текста с использованием тезауруса семантических полей

Одним из возможных решений поставленной задачи может стать включение в состав системы контроля знаний тезауруса семантических полей [4] и использование его при тестировании для автоматизированного построения концептуальной модели текста на естественном языке. Результат тестирования при этом определяется на основе сравнения полученной модели с концептуальными моделями предметной области или эталона ответа. Эталонная модель также может быть получена по неформализованному естественно-языковому тексту всего курса либо по тексту одного наиболее полного правильного ответа на конкретный вопрос. Таким образом, достоинством такого подхода является то, что эталонная модель строится автоматически, как и модель ответа тестируемого, что уменьшит трудозатраты автора. Эксперту по данной предметной области не обязательно разбираться в методах построения модели, и не надо перечислять все варианты ответов, достаточно дать один расширенный ответ на естественном языке. Несколько ответов может понадобиться ввести в том случае, если ответы на вопрос могут быть разными с концептуальной точки зрения. Такая ситуация может возникнуть в том случае, например, если проверяется знание понятия, которое имеет несколько верных толкований, которые могут использоваться в смежных областях знания.

Можно предложить несколько вариантов построения концептуальных моделей по тексту ответа, построенных с применением тезауруса семантических полей. В таблице 1 рассмотрены методы построения и анализа соответствия данных моделей.

Семантический объем понятия представляет собой набор семантических множителей, соответствующих содержанию этого понятия.

Понятие может быть представлено как словом, так и словосочетанием.

Методы и алгоритмы построения семантического объема и кодирования слов в семантические множители рассмотрены в [5].

5. Структура тезауруса

Под тезаурусом семантических полей не имеется в виду фиксированный дескрипторный тезаурус, аналогичный применяемым в информационно-поисковых системах или в системах автоматического индексирования. Для возможности использования тезауруса семантических полей в системе контроля знаний к нему предъявляются следующие требования:

- тезаурус семантических полей должен содержать информацию, необходимую для построения концептуальных моделей текста;

- для выполнения требования упрощения наполнения системы контроля знаний должны существовать методы автоматизированного построения тезауруса семантических полей.

Для реализации предложенных методов проверки правильности ответов, тезаурус семантических полей должен содержать информацию следующего вида (схема объектов и отношений, фиксируемых в тезаурусе, представлена на рисунке 1):

1. объекты:
 - перечень слов представляющий собой словарь тезауруса (СТ);
 - словоформы, соответствующие этим словам, которые встречаются в обрабатываемых текстах представляющий собой словник тезауруса (СЛТ);
 - определения слов, входящих в СТ в текстовой форме (О);
 - перечень семантических множителей – словарь множителей тезауруса (СМТ);
 - словарь устойчивых терминов (СТТ).
2. отношения, фиксируемые в тезаурусе:

Таблица 1

Методы построения концептуальных моделей текста

Наименование метода	Построение модели	Анализ соответствия моделей
1. Метод множества	По тексту строится семантический объем	Находится пересечение семантических объемов ответа и эталона, при наличии в семантическом объеме ответа более 50% семантических множителей, входящих в семантический объем эталона, ответ считается удовлетворительным.
2. Метод линейного множества	По тексту строится семантический объем с приписыванием семантическим множителям последовательности встречаемости слова в тексте.	Рассчитывается коэффициент корреляции, учитывающий, кроме вхождения в семантический объемы семантических множителей, и порядок следования слов.
3. Метод подстановки	Для каждого слова ответа ищется корреляционный заменитель (попадающий в то же парадигматическое семантическое поле) среди слов эталона, или наоборот.	Вводятся уровни соответствия моделей – если а) найдены все заменители; б) найдено более 50% заменителей; в) найдены все имеющие низкую частоту и высокую значимость в данной предметной области слова, - то ответ верен на определенном уровне.
4. Метод подстановки с учетом типа связи в предложении	При построении модели для каждого заменителя указывается тип связи с другими словами в ответе и эталоне.	Для анализа вводится корреляционный коэффициент, учитывающий, кроме проверки наличия заменителя, и совпадение связи между словами.

- связи между СТ и СлТ, отражающие, какая словоформа определенного слова появилась в обрабатываемых текстах (Отношение "слово-словоформа" - ОССл)
- словарь словоупотреблений (СС);
- связи между отношениями ОССл, отражающие наличие синтагматической связи между словами, в определенных формах, встретившиеся в текстах (отношение "ОСС") - словарь синтагматических ссылок (ССС);
- связь между словом и определением, показывающая наличие дефиниции данного слова, заполняющийся из толковых и терминологических словарей
- словарь дефиниций (СД);

При такой организации тезауруса, синтагматические связи фиксируются в словаре, а связи между словами и кодами, и парадигматические связи между словами могут быть получены алгоритмически, поэтому нецелесообразно хранить их в тезаурусе постоянно [6,7]. Алгоритмическое получение данных связей по необходимости имеет преимущество возможной модификации критериев выделения связей в зависимости от цели их получения (например, изменение частотных критериев при получении парадигматического поля, уровня компонентного анализа при построении семантического объема), что делает тезаурус более универсальным, расширяя сферу его применения.

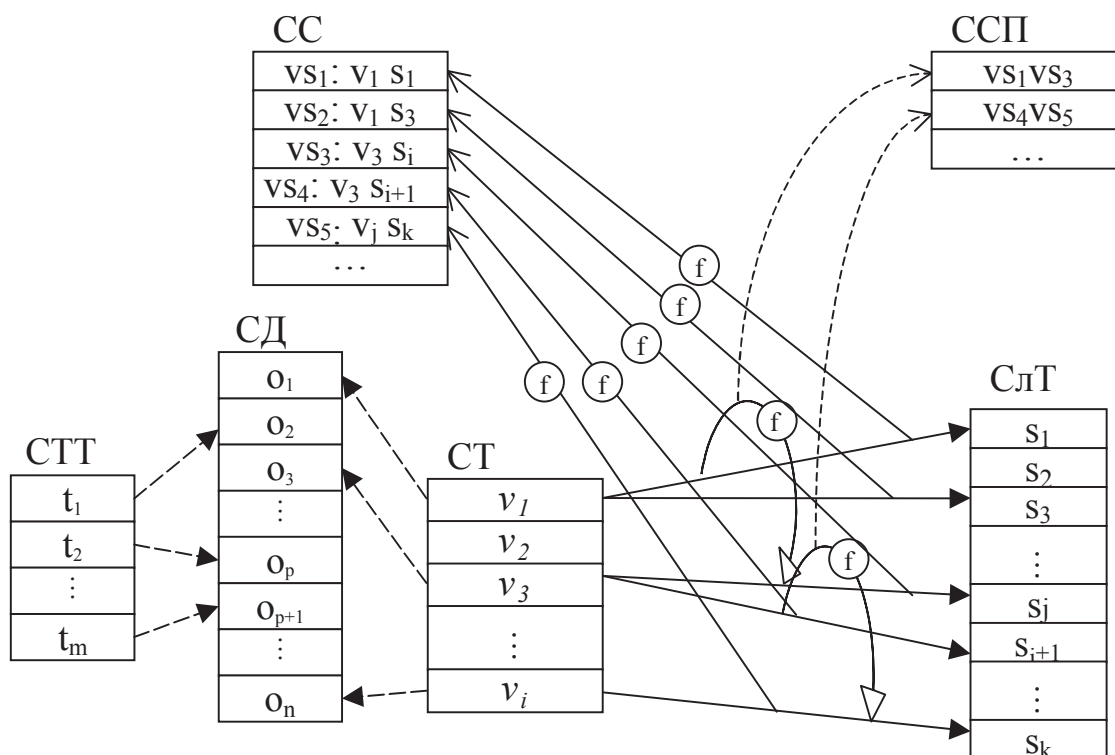


Рис. 1. Схема объектов и отношений в тезаурусе

6. Алгоритм наполнения тезауруса

На вход данного алгоритма могут поступать тексты на естественном языке, включая и тексты по тематике тезауруса, и обрабатываемые дефиниции. На выходе - тезаурус, изменивший свое состояние. t - анализируемый текст,

1-й шаг Разбиение текста t на массив входящих в него словоформ. $i=1$.

2-й шаг Получаем по словоформе S_i код. Находим в СлТ слово, которому соответствует данный код.

2.1 Если слова нет - перейти к шагу 3.

2.2 Если слово есть, перейти к шагу 4.

2.3 Если слов, соответствующих данному коду несколько, то применить алгоритм разрешения неоднозначности, определить, к какому слову относится данная словоформа, перейти к шагу 4.

3-й шаг Найти в словаре дефиниций определение, код слова которого соответствует анализируемому

коду. Если слово найдено, включаем данное слово в СТ. Если слово не найдено, то запросить у источника данных (внешний словарь или пользователь) информацию о слове и его определении по словоформе.

4-й шаг Если в словаре СлТ нет словоформы S_i , то включить S_i в СлТ;

5-й шаг Занести запись в словарь словоупотреблений, СС имеющий формат: код_словаря-код_словника-частота. Если запись о словоупотреблении (без частоты) отсутствует в СС, то вносим такую запись с частотой 1, иначе увеличиваем частоту нужной записи на 1.

6-й шаг Определить синтагматические связи в окружении слова, занести данные в таблицу синтагматических ссылок, имеющую формат: код_словоупотребления1-код_словоупотребления2-связь-частота. Если запись об определенной связи (без частоты) отсутствует в таблице синтагматических ссылок, то вносим такую запись с частотой 1, иначе увеличиваем частоту нужной записи на 1. $i=i+1$, переходим к шагу 2.

Предложенные схема и алгоритм построения тезауруса позволяют вносить в него любые типы синтагматических связей, дифференцируя их. В зависимости от целей применения тезауруса, можно выбирать различные их типы. Разные исследователи различали разные виды семантических связей [7,8], однако данная структура тезауруса позволяет сохранять и получать все основные виды синтагматических связей, для их нахождения можно использовать дистрибутивный и статистический анализы текстов выбранной тематики [9].

Для определения используемого в тексте актуализированного значения полисемантического слова можно использовать анализ семантических объемов значений слов контекста. Характерным для правильных значений слов, дающих оптимальное осмысление предложения, является наличие ряда общих семантических множителей [10]. Используемое значение полисемантического слова может определяться по максимальной повторяемости семантических множителей в пределах предложения.

Описанные методы позволяют выполнить автоматизированное построение тезауруса семантических полей, используя информацию толковых и терминологических словарей и естественно-языковые тексты выбранной предметной области.

7. Выводы

Предложенные методы позволяют решить некоторые проблемы повышения эффективности тестирования теоретических знаний в системах контроля знаний за счет уменьшения трудозатрат автора при наполнении системы. Для эффективного применения предложенных методов необходимо:

- совершенствование методов построения семантических полей, в частности комплексного поля;
- разработка эффективных механизмов анализа соответствия концептуальных моделей.

Разработанная структура тезауруса позволит осуществить тестирование не только предложенным методом (построением и анализом концептуальных моделей текста), но и применить метод анализа правильности ответа по ключевым словам, повысив его эффективность за счет автоматизированного построения списка ключевых слов. При этом в качестве ключевых слов могут быть использованы парадигматические поля слов, входящих в эталон ответа. Парадигматическое поле слова можно в явном виде в тезаурусе не сохранять, а при необходимости получать на основе сравнения семантических объемов по наличию у слов общих множителей. Для повышения значимости семантической связи в поле можно использовать критерий частотности семантического множителя. Так как более значимые множители будут иметь более низкую частоту появления, чем множители широкой семантики, следовательно, может быть установлена

наибольшая частота множителя, используемого для включения слова в поле. Таким образом может регулироваться ширина парадигматического поля и, следовательно, размер списка ключевых слов, что может быть использовано для градации правильности ответа.

Метод построения концептуальных моделей текста может быть также использован в документографических информационно-поисковых системах. При этом концептуальная модель первого вида может быть использована для предварительной проверки релевантности документа запросу, а модель второго вида – для более точной оценки.

Литература

1. Романов А. Н., Торощов В. С., Григорович Д. Б. Технология дистанционного обучения в системе заочного экономического образования. – М.: ЮНИТИ-ДАНА, 2000. – 303 с.
2. Автоматизированные обучающие системы на базе ЭВМ / [А. Ф. Чернявский, А. М. Мухарский, А. В. Орлов и др.] Под ред. А. Ф. Чернявского. – Мн.: Изд-во БГУ, 1980. - 176 с.
3. Бухараев Р. Г., Сулейманов Д. Ш. Семантический анализ в вопросно-ответных системах. – Казань: Изд-во КГУ, 1990. – 123 с.
4. Павлов П. Ф., Черепанова Ю. Ю., Шубин И. Ю. О возможности построения тезауруса семантических полей и его применения в информационных системах // Вісник ХДПУ. Збірка наукових праць. – Вип. 108. – Харьков: ХДПУ, 2000. – С. 41-46.
5. Черепанова Ю.Ю. Методы и алгоритмы построения семантического объема слова // Восточно-европейский журнал передовых технологий. Информационные технологии. 4/2(34)2008. С.21-25.
6. Черепанова Ю.Ю. О теоретико-множественном и теоретико-категорном подходах к моделированию семантических полей // Проблемы бионики: Всеукр. межвед. науч.-техн. сб. – 2001. – Вып 54. – С.75-78.
7. Павлов П.Ф., Бабина О.И., Черепанова Ю.Ю. О проблеме автоматизированного выявления семантических полей. // Вісник ХДПУ. Збірка наукових праць. – Вип. 42. – Харьков: ХДПУ, 1999. – С.81-85.
8. Полевые структуры в системе языка / З. Д. Попова, И. А. Стернин, Е. И. Беляева и др. – Воронеж: Изд-во ВГУ, 1989. – 196 с.
9. Арнольд И. В. Основы научных исследований в лингвистике: Учебное пособие. – М.: Высшая школа, 1991. – 140 с.
10. Апресян Ю. Д. Избранные труды. Том 1. Лексическая семантика. – М.: Восточная литература, 1995. – 472 с.