Козлов О.А., д.п.н., проф.; Шаров А.Г., г.Москва

ПРОБЛЕМЫ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТА В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ КОНТРОЛЯ ЗНАНИЙ

В настоящее время применение информационных технологий и информационной техники в школьном образовательном процессе становится рядовым явлением. Совокупность задач, а также естественная эволюция информационного сообщества, сделали актуальными исследования в области автоматизации образования в целом, и совершенствование методов автоматизированного контроля знаний в частности.

Наибольшую сложность для практической реализации в автоматизированных системах контроля знаний представляет оценивание ответов на задания открытого типа, ответы на которые должны быть даны в виде произвольного текста. К таким заданиям относятся, например, задания на знание тезауруса предметных областей, требующие дать определение некоторого термина изучаемой предметной области, то есть его формулировку.

Разработанные к настоящему времени методы автоматизированного контроля ответов на такие задания выделяются в несколько основных направлений, среди которых следует отметить следующие:

– Каталогизация и ранжирование возможных вариантов ответов.

– Семантический анализ ответов на основе ключевых слов и словосочетаний.

Суть методов каталогизации и ранжирования сводится к следующему. При составлении задания автор должен указать единственный правильный ответ, либо привести, по его мнению, полный перечень возможных ответов, корректных в большей или меньшей степени. Как правило, при этом каждому варианту ответа ставится в соответствие некоторое количество баллов. Анализ выполняется на основе совпадения тестируемого ответа с одним из перечня, составленного автором.

Основным достоинством таких методов анализа является простота их реализации в автоматизированных системах. К недостаткам можно отнести значительную трудоемкость при составлении заданий, большие массивы хранимой информации, необходимость составления перечня ответов для ***каждого*** задания.

Методы семантического анализа представляют относительно новый вид решения задач анализа произвольного текста. Эти методы в основном применяются в поисковых системах в сети Интернет, однако могут быть адаптированы к задачам автоматизации контроля знаний. Их суть сводится к описанию ключевых слов и словосочетаний, которые могут быть использованы в ответах. Анализ ответа выполняется на основе выявления в

ответе словоформ ключевых слов, указанных автором при составлении задания.

К достоинствам методов семантического анализа можно отнести меньшую трудоемкость при составлении заданий, по сравнению с методами каталогизации, а также устойчивость методов к грамматическим неточностям в ответах. Основным недостатком таких методов является сложность их практической реализации и недостаточная адаптация механизмов их реализации к задачам образовательного назначения.

Нами было проведено исследование, выявившее низкий уровень эффективности существующих в настоящее время автоматизированных систем контроля знаний в части анализа ответов на задания открытого типа. Это определило необходимость разработки нового подхода к данной проблеме.

Проведенное нами исследование направлено на оптимизацию автоматизированного анализа ответов на задания открытого типа. Исследовалась проблема анализа ответов на задания, требующие ввести с клавиатуры определения некоторых терминов изучаемой предметной области. Дополнительным условием разработки являлось приблизить адекватность разрабатываемых методов к анализу ответа при устном опросе.

За основу нами было принято утверждение о том, что опрашиваемые не всегда в точности повторяют определение термина, данное преподавателем, но излагают ***смысл*** определения, заменяя некоторые его слова близкими им по значению терминами, то есть синонимами, в соответствии с принятыми правилами грамматического построения предложений.

Исходя из вышесказанного, в качестве основы решения нами предложено использовать синонимические ряды. В настоящее время широко известны словари синонимов Александровой, Шведова и других авторов, что позволило использовать их в достижении целей нашего исследования. Также, каждая предметная область имеет свой собственный словарь и тезаурус. В практической реализации автоматизированной системы контроля знаний на основе излагаемой методики словари синонимов и тезаурусы предметных областей единожды должны быть занесены в хранилище системы.

Суть разработанной нами методики состоит в нахождении релевантности тестируемого ответа относительно ответа автора путем анализа этих ответов с применением синонимических рядов.

При составлении задания автор должен ввести эталонный ответ. Слушатель вводит тестируемый ответ.

Представим совокупность синонимических рядов в виде таблицы размером $\boldsymbol{N}\times\boldsymbol{N}$, где $\boldsymbol{N}$ – количество слов в словаре изучаемой предметной области. Каждая строка таблицы соответствует отдельному термину, столбцы и строки с одинаковыми порядковыми номерами соответствуют одному и тому же термину. В ячейки, соответствующие паре различных терминов, имеющих синонимическую связь друг с другом, вписаны единицы; в остальные ячейки вписаны нули. Если вдоль столбцов таблицы расположить термины, встречающиеся в эталонном ответе, а вдоль строк –

термины, встречающиеся в тестируемом ответе, то совокупность ячеек, в которых вписаны единицы, соответствующие одновременно терминам, содержащимся в обоих ответах, укажет релевантность тестируемого ответа относительно эталонного ответа.

Данный метод позволяет определить качественный показатель релевантности тестируемого ответа. Для вычисления количественного показателя возможно использование нескольких принципиально различных математических методов.

В основе одного из них лежит ранжирование синонимов по семантической удаленности в пределах каждого синонимического ряда. Иными словами, для каждого термина синонимического ряда назначается числовой показатель синонимии с термином, относительно которого составлен ряд. Значение числового показателя варьируется от 0 до 1. «1» означает абсолютную идентичность слов. Чем меньше значение показателя, тем меньше синонимическая связь между словами. Применение этого метода подразумевает, что единичные значения в таблице синонимии заменены числовыми показателями от 0 до 1, связывающими термины эталонного и тестируемого ответов. Дальнейшее решение предполагает выработку математической формулы для вычисления числового показателя релевантности ответа. Получение данной формулы, по нашему мнению, возможно лишь эмпирическим путем, так как известные теоретические методы поверхностно затрагивают рассмотрение подобных задач.

Другой метод вычисления показателя релевантности основан на теории нейроподобных вычислительных структур. В соответствии с ним необходимо сопоставить каждому тестируемому ответу его образ-вектор, отображающий эталонный ответ в некотором пространстве, таким образом, чтобы вектора сходных ответов были близки. Эта задача относится к задачам классификации, и для ее решения можно использовать сети динамической ассоциативной памяти и модификации сети Хемминга.

Независимо от выбранного метода вычисления показателя релевантности, необходимо получить его количественную характеристику. Эта характеристика, обозначающая результаты тестирования, пригодна для выработки автоматических корректирующих воздействий по адаптации курсов автоматизированного и дистанционного обучения, в том числе и в разрезе индивидуального подхода к обучаемым, что позволит повысить качество автоматизированных систем, применяемых в дистанционном образовании.