УДК 378.168:004

СОСТОЯНИЕ И УРОВЕНЬ РАЗРАБОТОК СИСТЕМ АВТОМАТИЧЕСКОЙ ОЦЕНКИ СВОБОДНЫХ ОТВЕТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Мишунин О.Б., Савинов А.П., Фирстов Д.И.

ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», Томск, e-mail: savinov@tpu.ru

В настоящей статье делается обзор зарубежных и отечественных работ, посвящённых оценке свободноконструируемых ответов на естественном языке. С целью классификации обозреваемые системы были разделены на группы, которые отличаются друг от друга принципом представления и анализа текста эталона и ответа обучаемого: 1. Создание шаблонов, представляющих в усечённом виде ключевые слова в ответном тексте. 2. Представление текста в виде неупорядоченного множества ключевых слов (т. н. «мешок слов» — «bag of words»). 3. Учёт синтаксических или семантических связей слов в предложении. Показано, что системы с первым типом оценки ответов открытого типа дают качественные и предсказуемые результаты при условии учёта в шаблоне всех возможных формулировок правильного ответа. Однако выполнение этого условия приводит к перегрузке преподавателя из-за необходимости осуществления большого объёма подготовительных работ. Для систем второго типа характерна проблема, связанная с моделью «мешок слов»: можно сформулировать ответ из ключевых слов, соответствующих эталону, который получит высокую оценку, будучи при этом совершенно бессмысленным. Показано, что системы контроля, относящиеся к третьей группе оценки свободных ответов, отстают по своему качеству от других систем. Указаны проблемы, которые необходимо решить, чтобы они заняли лидирующее положение относительно других рассмотренных групп.

Ключевые слова: автоматическая оценка ответов на естественном языке, интеллектуальные обучающие системы, компьютерная лингвистика, электронное обучение, семантика

STATE AND LEVEL OF THE AUTOMATIC FREE-TEXT ANSWER GRADING SYSTEMS DEVELOPMENT

Mishunin O.B., Savinov A.P., Firstov D.I.

National Research Tomsk Polytechnic University, Tomsk, e-mail: savinov@tpu.ru

In this article we review intelligent tutoring systems with automatic free-text answer grading. Free-text answer tasks is one of the most natural knowledge assessment forms familiar to both teachers and students. Based on how student answer is compared to reference answer we have classified reviewed systems into groups that use following approaches: 1. Template-based approach. 2. «Bag of words» approach. 3. Use of relations between words in text and their roles and functions in it. It is shown that the first group systems have predictable high quality results providing the answer template covers all the possible variants of correct student answer. This condition may require a large amount of work from teachers developing the course. Systems using «bag of words» method only take into account the number of key words presented in both student answer and reference answer, which makes it possible to write a meaningless answer using the right words and get high grade for it. Systems of the third group are promising but currently fall behind the other two groups in their efficiency. We have shown some of the problems that need to be solved in order for these systems to be as efficient as others.

Keywords: automatic assessment of free-text answers, intelligent tutoring systems, computational linguistics, e-learning, semantics

Высокая эффективность процесса обучения достигается не только применением современных дидактических и технических приёмов, но и характером организации обратной связи с обучаемым. Последнее обеспечивается в интеллектуальных обучающих системах контролем знаний обучаемого, осуществляемого автоматическими тестирующими подсистемами. Тестовые задания по форме могут быть открытыми или закрытыми. Открытые задания требуют от испытуемого свободно конструируемого ответа, а закрытые предлагают выбрать правильный ответ из нескольких предложенных вариантов.

Основным негативным моментом, присущим закрытым вопросам, является то, что такие вопросы требуют от студента узнавания правильного ответа, а не конструирования. Соответственно, с помощью таких вопросов легко проверять память студента, но не логику его мышления или креативные способности. Другая проблема закрытых вопросов — это необходимость для преподавателя придумывать правдоподобные неправильные ответы, что в разы повышает сложность подготовки тестового материала.

Тест со свободно-конструируемыми ответами является наиболее естественной и привычной формой контроля знаний. Тестируемый получает вопрос и формулирует ответ на естественном языке, а система оценивает этот ответ. Основной задачей интеллектуальной системы, анализирующей свободный ответ студента, является выявление смысла, заложенного в нём. Этот смысл сравнивается с заведомо правильным, извлекаемым из эталонного ответа. Правиль-

ность и глубина ответа определяется системой по степени его смысловой близости к эталонному ответу. Работы по созданию программных средств, направленных на извлечение из произвольного текста смысла, инвариантного к грамматическому представлению, находятся в настоящее время в стадии, далёкой от завершения. Поэтому на данном этапе исследований можно осуществлять только приблизительную оценку свободных ответов путём сравнения отдельных ключевых слов и фраз, входящих в текст эталонного ответа и ответа студента.

В статье делается обзор работ, описывающих зарубежные и отечественные системы, в которых используется оценка свободных ответов на естественном языке, и анализируются методы, на базе которых возможно создание таких систем. Обозреваемые системы были разделены на группы, отличающиеся друг от друга принципом представления и анализа текста эталона и ответа обучаемого:

- 1. Создание шаблонов-масок.
- 2. Представление текста в виде неупорядоченного множества слов.
- 3. Учёт ролевых функций слов в предложении и их связей между собой.

Системы на основе шаблонов-масок

Системы, относящиеся к данной группе, рассматривают ответный текст как упорядоченный набор символов. Преподаватель подготавливает шаблон (маску) на формальном языке, учитывающий возможные перефразировки правильного ответа. В шаблоне закладываются логические ограничения на входящие в ответ подстроки. Преподаватель может указать обязательные подстроки, присутствие которых необходимо в правильном ответе. Можно задать список альтернативных вариантов, и тогда система будет искать в ответе только одну подстроку из этого списка. Можно также задать список подстрок, которых в правильном ответе быть не должно. Все эти условия могут быть скомбинированы в одном шаблоне (см. пример ниже). В качестве подстрок обычно указываются ключевые слова, в которых оставляется только та часть, которая не меняется при словообразовании. Для того чтобы система могла проверять текст с возможными орфографическими ошибками, часть букв заменяется на «знак безразличия» - специальный символ, вместо которого в ответе может стоять любая буква.

Первая система, которая реализует описанный принцип, называется КИТ (расшифровывается как «компьютерный интеллектуальный тьютор») [6, 9]. Эта система была разработана в Международном банковском институте (Санкт-Петербург). Она

предназначена для ведения обучающих диалогов со студентами на естественном языке. Рассмотрение диалоговой составляющей выходит за рамки нашего обзора, поэтому сосредоточимся на том, как система определяет правильность ответа. Критерии проверки ответа задаются в виде шаблонов на языке ФЛОД (функция логическая основных дескрипторов) [2, 10].

Чтобы проиллюстрировать работу данной системы, а также общий принцип, по которому работают остальные системы, описанные в этом разделе, составим шаблон на языке ФЛОД. В качестве примера будем использовать вопрос из курса «Экономика предприятия»: В чём отличие понятий «сырьё» и «материалы»? Эталонный ответ на этот вопрос: Сырьё – это продукция добывающих отраслей, а материалы – это сырьё, прошедшее первичную обработку.

Формализуем эталонный ответ из примера с помощью языка ФЛОД:

сырь & ((продук & добыв & (отрасл ! произв)) ! (природ & р*сурс)) & материал & сырь & обработ & $\overline{\text{не}}$

В данном примере & — это символ логической конъюнкции (И), ! — дизьюнкции (ИЛИ), * — «знак безразличия». Черта над текстом означает, что данная последовательность символов должна отсутствовать в правильном ответе. Следует отметить, что мы несколько расширили формулировку эталона, добавив вариативности.

Похожий подход реализован в системе WebLAS [12, 15], разработанной в Калифорнийском университете (Лос-Анджелес). Здесь также используется формализованный язык описания шаблонов, но с другим синтаксисом. Ключевое отличие системы WebLAS заключается в том, что построение шаблонов автоматизировано. Преподаватель вводит эталонный ответ на естественном языке и отмечает ключевые слова. Система подбирает синонимы для отмеченных слов. Преподаватель обрезает ключевые слова и подобранные синонимы, оставляя только те части, которые не изменяются при склонении и словообразовании. Шаблон генерируется на основе предоставленной информации автоматически.

В работе Д.А. Мерзлякова из Пермского государственного национального исследовательского университета [4] представлен подход к проверке свободных ответов на основе регулярных выражений. Для упрощения написания шаблонов предлагается использование метарегулярных выражений — выражений на разработанном автором формализованном языке, позволяющем описывать типовые структуры, характерные определённым видам текста (см. при-

мер шаблона для определений в указанной статье). Использование метарегулярных выражений позволяет единожды задать шаблоны для описания возможных перефразировок одной языковой структуры (например, определения). Подставляя в этот шаблон усечённые ключевые слова, можно получать регулярные выражения сразу для всех заданных ранее перефразировок.

Рассмотрим проблемы описанной группы систем. Теоретически, использование созданных вручную шаблонов для автоматической проверки свободных ответов должно давать качественные и предсказуемые результаты при условии учёта в шаблоне всех возможных формулировок правильного ответа. На практике, выполнение этого условия приводит к перегрузке преподавателя из-за необходимости осуществления большого объёма подготовительных работ. При этом учесть все варианты возможных ответов студентов не представляется возможным.

Другая проблема данного подхода — вероятность совпадения части неправильного ответа с шаблоном. Это происходит из-за того, что преподаватели стремятся писать шаблоны как можно более кратко, включая в них только необходимые ключевые слова. Студент же может дать длинный неправильный ответ, который, тем не менее, будет содержать все необходимые слова из шаблона.

Системы, основанные на модели «мешок слов»

Системы данной группы рассматривают ответный текст как неупорядоченный набор слов, *п*-грамм (последовательностей из *п* идущих подряд слов) или устойчивых выражений (обычно не более 2–3 слов). За рубежом эта модель получила название «bag of words» («мешок слов»). Преподаватель вводит эталонные ответы в виде обычного текста. Оценка ответа студента пропорциональна числу слов в этом ответе, совпавших со словами из эталона.

Проиллюстрируем суть этого подхода на примере. Предположим, что эталонный ответ преподаватель задал в таком виде: Сырьё — это продукция добывающих отраслей, а материал — это сырьё, прошедшее первичную обработку. Студент дал такой ответ: Сырьё — это природный продукт, а материал изготавливается из сырья. Представив два текста в виде множеств начальных форм и определив их пересечение, мы получим множество из пяти элементов: сырьё, это, а, материал, сырьё. Нормировав количество элементов во множестве на количество элементов в эталоне, получим оценку ответа студента. Она будет очень не-

высока (~ 0.42), несмотря на то, что студент ответил, по сути, правильно.

Чтобы избежать занижения оценки, можно пойти разными путями. Первый способ – использовать множество эталонных ответов. Например, существует большое количество систем, в которых используется машинное обучение по корпусу оценённых ответов с применением модели «мешок слов» [11, 24–27, 34, 39, 46]. Здесь мы не будем рассматривать конкретные системы, так как они отличаются лишь используемыми методами машинного обучения, обзор которых выходит за рамки нашей статьи.

В системе Atenea [37, 38] (Мадридский автономный университет, Испания) для оценки свободных ответов был применён метод BLEU (bilingual evaluation understudy). В методе BLEU оценка ответа пропорциональна проценту *п*-грамм ответа студента, содержащихся в корпусе эталонных ответов. Также учитывается длина ответа, чтобы неполные ответы не получили завышенную оценку.

Второй способ – использование лингвистических баз знаний. Так, в статье [41] представлена система ASAGS, разработанная в Национальном институте технологий (Тируччираппалли, Индия). В этой системе используются тезаурус WordNet. С его помощью снимается синонимия, расшифровываются аббревиатуры, дериваты заменяются исходным словом. После предварительной обработки слов в ответных текстах оценка вычисляется как процент совпадения слов в эталоне и ответе студента.

В работе [5] представлена система, разработанная в Томском политехническом университете, использующая модель «мешок слов» в сочетании с лингвистическими базами знаний на основе Википедии и тезауруса Рутез. По базам знаний рассчитывается семантическая близость (значение от 0 до 1) между найденными в тексте терминами. Под терминами понимаются как отдельные слова, так и устойчивые многословные именные конструкции. Результирующая оценка пропорциональна сумме оценок семантической близости между терминами эталона и студенческого ответа.

В качестве базы знаний для оценки семантической близости понятий может использоваться онтология предметной области. Так, в статьях [16, 17] описывается система OeLE (Университет Мурсии, Испания), в которой эталонный ответ представляется в виде набора элементов онтологии: понятий, атрибутов и отношений. Такие же элементы полуавтоматически извлекаются из ответа студента. Между извлечёнными из студенческого ответа элементами онто-

логии и заданными преподавателям эталонными онтологическими элементами рассчитывается семантическая близость. Сумма получившихся оценок формирует конечную оценку студенческого ответа.

Третий способ устранения занижения оценки - использование векторных моделей представления текста. Идея подхода в том, что смысл слова можно выразить контекстом, в котором оно часто встречается. Векторные модели требуют предварительного обучения на корпусе текстов. Корпус разбивается на документы – семантически единые отрезки текстов (статьи, главы, абзацы и т.п.). Смысл слова описывается многомерным вектором в «семантическом пространстве». Число измерений вектора равно числу документов в корпусе. Значения в каждой размерности – количество повторений слова в соответствующем документе (или другая мера, прямо пропорциональная числу повторений слова в документе, например, TF-IDF). Семантическая близость между двумя отдельными словами рассчитывается как косинус угла между представляющими их векторами. Текст также может быть представлен в виде вектора в «семантическом пространстве». Близость между двумя текстами также определяется через косинус угла между представляющими их векторами.

Первое известное нам применение векторного подхода к оценке свободных ответов опубликовано в статье [29]. В этой статье описывался метод латентно-семантического анализа (ЛСА). В ЛСА по корпусу документов рассчитывается термдокументная матрица, в которой строки соответствуют словам, столбцы - документам, а в ячейках находится значение TF-IDF для данного слова в данном документе. После этого к матрице применяется сингулярное разложение, в результате чего снижается её размерность. Положительными эффектами снижения размерности являются удаление незначительных слов и объединение синонимов. После этого близость ответов рассчитывается путём представления их в виде векторов и вычисления косинуса угла между ними.

На основе ЛСА построено множество систем автоматической оценки ответов. Так, в статьях [19–22] описывается интеллектуальная диалоговая система обучения AutoTutor (Университет Мемфиса, США). Эталонные ответы в системе задаются в виде наборов мелких тезисов. Все тезисы должны быть произнесены студентом во время общения с машиной. Наличие тезиса в ответе студента определяется с помощью ЛСА.

Метод ЛСА хорошо комбинируется с другими методами обработки естественно-языковых текстов. Так, в статье [36] показано, что можно добиться повышения качества проверки с помощью ЛСА, предварительно обрабатывая слова в тексте: приводя их к начальной форме, удаляя служебные слова и разрешая лексическую омонимию. В работе [23] комбинировали ЛСА и оценку по совпадению n-грамм. Ответ студента оценивался отдельно с помощью двух методов, а в качестве конечной оценки бралось среднее арифметическое. Такая комбинация также привела к повышению качества оценки. Авторы статьи [28] пытались добавить в модель «мешок слов» синтаксическую информацию. Для этого каждое слово в ЛСА было дополнено информацией о частях речи слов, окружающих данное. Это привело к снижению качества проверки.

Другой векторный подход — ESA (explicit semantic analysis) — был описан в статье [18], а в статье [35] применён для оценки свободных ответов. Этот метод использует в качестве корпуса Википедию, а одним документов считается одна статья. ESA показал лучшую корреляцию с оценками преподавателей по сравнению с методом ЛСА, использующим тот же корпус.

Рассмотрим проблемы описанного класса систем. Системы, построенные на базе данной модели, не подходят для оценки ответов, в которых важен порядок следования текстовых фрагментов (вопросы про алгоритмы или хронологические последовательности) или важна структура (вопросы на классификацию или достоинства и недостатки некоторого явления).

Анализ ограничений модели «мешок слов» наводят на мысль о важности учёта связей между словами в предложении при оценке свободных ответов.

Вполне очевидно, что, зная особенности работы алгоритма проверки, основанного на модели «мешок слов», можно сформулировать из входящих в эталон слов такой ответ, который получит высокую оценку, будучи при этом совершенно бессмысленным. Исходя из этого, существует ненулевая вероятность того, что студент даст такой же бессмысленный ответ, а система зачтёт его как правильный.

Системы, учитывающие ролевые функции слов в предложении и их взаимную связь

Системы данной группы рассматривают текст как набор связанных слов. Связи извлекаются из текста с помощью синтаксического или семантического анализа. В системах автоматической проверки обычно

используется два типа связей: пара слов, связанных атрибутивной или обстоятельственной связью, и тройка «субъект – предикат – объект». Для иллюстрации этого представления обратимся к нашему примеру: Сырьё – это продукция добывающих отраслей, а материал – это сырьё, прошедшее первичную обработку. Из этого примера могут быть извлечены такие связи:

сырьё есть продукция продукция отрасли материал есть сырьё отрасль добывающая сырьё прошло обработку обработка первичная

Из ответа студента также извлекаются связи. Далее они сравниваются с эталонными связями с учётом синонимичных замен. Конечная оценка пропорциональна количеству совпавших эталонных связей и связей, извлечённых из ответа студента.

В работах [29, 36] представлена система Auto-marking, разработанная в Великобритании в организации UCLES Кембриджского университета. Эта система делает морфологический и поверхностный синтаксический анализ студенческих ответов, в результате чего выделяются именные и глагольные группы. Для проверки ответов преподаватели конструируют шаблоны на формальном языке, в которых описываются ограничения на эти группы. Система с аналогичным подходом была разработана в университете Sir Syed University of Engineering and Technology в городе Карачи в Пакистане. Она называется IndusMarker [42-44]. В работах, посвящённых этой системе, описывается язык QAML, основанный на XML. Его возможности схожи с возможностями языка из предыдущей системы.

В работах [32, 33] описана система AutoMark, разработанная в компании Intelligent Assessment Technologies Ltd. В этой системе шаблон ответа представляется в виде синтаксического дерева. В узлах дерева преподаватель указывает возможные перефразирования для данного слова. При оценке ответа проводится его синтаксический анализ и строится дерево зависимостей. Затем это дерево сравнивается с эталонным. В случае совпадения деревьев ответ считается правильным.

В работах [3, 7, 8] описывается модель двухуровневого лингвистического процессора для анализа ответных текстов, разработанная Д.Ш. Сулеймановым (Казанский федеральный университет). Шаблон ответа в этой системе состоит из двух частей, используемых на разных уровнях анализа текста. В первой части шаблона строится таблица лексем с их ожидаемыми семантическими ролями (концептулами) в от-

ветном тексте. Во второй части шаблона записываются индивидуальные концептуальные грамматики (ИКГ). В терминологии Д.Ш. Сулейманова ИКГ – это схемы сочетания концептул в правильном ответе, зависящие от типа вопроса. Оценка ответа также выполняется в два этапа. На первом этапе в ответном тексте выискиваются лексемы, описанные в первой части шаблона ответа, и переводятся в соответствующие им роли (концептулы). На втором этапе последовательность концептул анализируется с привлечением ИКГ. На выходе двухуровневого лингвистического процессора формируется вектор ситуаций, отражающий правильность и полноту ответа.

В работе [1] рассказывается о програм-«Семантик-тест» (Тюменский государственный нефтегазовый университет), в которой используется представление эталонного текста в виде семантической сети. Оценка студенческого ответа выполняется следующим образом. Система сначала проводит синтаксический анализ ответного текста, в результате чего получается набор синтаксических отношений между словоформами. Затем проводится семантический анализ, на выходе которого получается семантическая сеть. В узлах семантической сети находятся именные группы, в рёбрах записываются предикаты. Семантические сети отдельных предложений объединяются по кореферентным именным группам. Затем сети эталонного ответа и ответа студента сравниваются с использованием тезауруса предметной области и правил перефразирования.

В статьях [13, 14] описывается прототип системы АТМ (Портсмутский университет, Великобритания), предназначенной для оценки вопросов по фактографическим дисциплинам. Эталонный ответ и ответ студента представляются в виде групп зависимостей, где синонимичные слова и фразы приводятся к единому виду с помощью встроенных в систему тезауруса предметной области и тезауруса фундаментальных понятий. Конечная оценка студенческого ответа пропорциональна количеству групп зависимостей, совпавших в ответе студента и эталонном ответе.

Более сложный подход реализован в системе C-rater (Educational Testing Service, Принстон, США) [30, 31, 45, 47]. Эталонный ответ в этой системе строится с помощью специализированного программного обеспечения. Он представляет собой множество ключевых аспектов правильного ответа, выраженных в форме простых предложений. Каждый аспект может иметь неограниченное количество формулировок.

В каждой формулировке преподаватель отмечает ключевые слова. При анализе студенческих ответов также выделяются тройки «субъект – предикат – объект». После этого разрешается местоименная кореферентность и слова в ответе заменяются на семантически близкие слова из эталонного ответа. Семантическая близость заранее рассчитывается автоматически по корпусу текстов, исходя из предположения, что слова, имеющие схожий смысл, имеют похожую сочетаемость в текстах. В результате получается нормализованная форма ответа, максимально близкая к эталону. После этого с помощью системы правил выделенные и обработанные тройки сравниваются с эталоном. При сравнении троек система не требует точного совпадения зависимых слов – достаточно, чтобы совпали те, которые преподаватель отметил, как ключевые.

Также существуют подходы, использующие методы машинного обучения в сочетании с синтаксическими признаками ответных текстов [26, 40]. Как отмечают авторы этих работ, за счёт большого обучающего корпуса повышается качество работы метода проверки.

Исследователи столкнулись с теми же проблемами, что и при использовании нелингвистических шаблонов. Отсутствие логического вывода приводит к тому, что за некоторые ответы студентам ставится заниженная оценка. Ориентированность систем на поиск правильной информации в ответе студента приводит к завышению оценки в случае, если ответ правилен лишь частично. Также отмечались проблемы, обусловленные с неустранимыми автоматически грамматическими ошибками в студенческих ответах.

Заключение

В рамках задачи построения интеллектуальных обучающих систем существует значительное количество работ, направленных на автоматизацию проверки свободных ответов на естественном языке. Анализ этих работ показывает следующее. На сегодняшний день отсутствуют системы, позволяющие непосредственно оценивать смысловую близость естественно-языковых ответов студентов к эталонному ответу. Существуют только системы, работа которых основана на принципе сличения параметров, косвенно отражающих контекстное содержание сравниваемых текстов. К их числу можно отнести:

• системы, основанные на использовании различного вида шаблонов (масок), обеспечивающих независимость представления ключевых слов в ответах от их лекси-

ческой формы и орфографических ошибок при написании;

- системы, представляющие текст в виде неупорядоченного множества слов, устойчивых выражений или *n*-грамм;
- системы, учитывающие порядок и связи слов в тексте, которые получаются с помощью синтаксического или семантического анализа.

Наиболее точную оценку свободных ответов студентов на естественном языке даёт сегодня группа систем на основе шаблонов. Несмотря на то, что они дают сейчас лучшие результаты, будущее, по нашему мнению, за направлением, осуществляющим семантико-синтаксический анализ предложений, определяя ролевые функции и контекстуальные связи слов в тексте ответов. Слабая их эффективность обусловлена наличием нерешённых проблем, которые определяют выходные характеристики лингвистического процессора системы контроля знаний. К их числу можно отнести следующие:

- 1. Необходимо автоматизировать трансформацию текстовой информации сложной грамматической конструкции в совокупность простых или сложноподчиненных предложений. Это позволит унифицировать синтаксическую и семантическую обработку произвольных текстов и, таким образом, решить проблему получения семантически эквивалентной информации, независимой от грамматической конструкции ответов студентов.
- 2. Не решена до конца проблема разрешения кореферентности, в том числе исключение местоимённой анафоры в тексте, в результате чего нарушается контекстуальная связь внутри и между предложениями.
- 3. Не решена проблема автоматического наполнения баз знаний на основе неструктурированного текста.
- 4. Остаётся актуальной проблема автоматизированного создания по исходному тексту предметной онтологии, которая является толковым (понятийным) словарём в машинном представлении, содержащим информацию обо всех понятиях (терминах), их свойствах, отношениях между ними и присущих им ограничениях.
- 5. Не решена проблема логического вывода на основе информации, содержащейся в естественно-языковом тексте, из-за чего существующие системы лишь сравнивают конкретные формулировки (допуская большую или меньшую их вариативность), а не оценивают правильность отражённых в ответе знаний.

Работы в этом направлении в настоящее время активно ведутся, и следует ожидать, что в недалёком будущем вышеперечисленные проблемы будут решены.

Список литературы

- 1. Бидуля Ю.В. Методы и алгоритмы смыслового описания контента в системах тестирования: автореф. дис. канд. филол. наук. Тюмень, 2011.-25 с.
- 2. Джалиашвили 3.О., Стригун А.И. Анализатор естественно-языковых ответов в АОС // Управляющие системы и машины. -1989. -№ 5. -С. 119-121.
- 3. Зуенко А.А. Алгебраическая модель вопросно-ответного диалога для построения индивидуальной траектории обучения // Труды Кольского научного центра РАН. -2014. Т. 24, № 5. С. 150–157.
- 4. Мерзляков Д.А. Генерация регулярных выражений для автоматизации проверки тестов открытого характера // Ноосфера. Общество. Человек. 2013. № 4.
- 5. Мишунин О.Б., Савинов А.П., Фирстов Д.И. Проблемы, возникающие в интеллектуальных обучающих системах при оценке ответов на естественном языке // Современные проблемы науки и образования. -2015. -№ 2.
- 6. Стригун А.И. Компьютерные интеллектуальные тьюторы // Образовательные технологии. 2014. № 4. С. 99–108.
- 7. Сулейманов Д.Ш. Двухуровневый лингвистический процессор ответных текстов на естественном языке // Сборник трудов Международной научно-технической конференции OSTIS-2011. 2011. С. 311–322.
- 8. Сулейманов Д.Ш., Аюпов М.М. Семантический анализ естественно-языковых текстов в вопросно-ответном режиме // Казань: Изд-во «Фэн» Академии наук РТ. 2011. С. 344–348.
- 9. Тазетдинов А.Д. Автоматный подход в построении компьютерных обучающих диалогов // Информационноуправляющие системы. -2008. -№ 4. -C. 42–48.
- 10. Тазетдинов А.Д., Тазетдинов Д.Р. Автоматизированные обучающие диалоги в естественно-языковой электронной образовательной среде // Международный журнал экспериментального образования. 2010. № 1. С. 101–103.
- 11. Alotaibi S.T., Mirza A.A. Hybrid approach for automatic short answer marking // Southwest decision sciences forty-third annual meeting. 2012.
- 12. Bachman L.F. et al. A reliable approach to automatic assessment of short answer free responses // Proceedings of the 19th international conference on Computational linguistics. $-2002.\,-T.\,2.$
- 13. Callear D., Jerrams-Smith J., Soh V. CAA of short non-MCQ answers // Proceedings of the 5th CAA Conference, Loughborough: Loughborough University. 2001.
- 14. Callear D., Jerrams-Smith J., Soh V. Bridging gaps in computerised assessment of texts // IEEE International Conference on Advanced Learning Technologies. 2001. P. 139–140.
- 15. Carr N.T., Xi X. Automated scoring of short-answer reading items: implications for constructs // Language Assessment Quarterly. -2010.-T.7, N_{$\!$ 0} 3. -P.205-218.
- 16. Castellanos-Nieves D. et al. Semantic Web Technologies for supporting learning assessment // Information Sciences. 2011. T. 181, N 9. P. 1517–1537.
- 17. Frutos-Morales F. et al. An extension of the OeLE platform for generating semantic feedback for students and teachers // Social and Behavioral Sciences. −2010. −T. 2, № 2. −P. 527–531.
- 18. Gabrilovich E., Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis // IJCAI International Joint Conference on Artificial Intelligence. —2007. P. 1606–1611.
- 19. Graesser A.C. et al. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue // IEEE Transactions on Education. -2005. -T. 48, N $\!\!\!_{2}$ 4. -P. 612–618.
- 20. Graesser A.C. et al. AutoTutor: A simulation of a human tutor // Cognitive Systems Research. 1999. T. 1, № 1. P. 35–51.
- 21. Graesser A.C. et al. Using latent semantic analysis to evaluate the contributions of students in AutoTutor // Interactive Learning Environments. 2000. T. 8, № 2. P. 129–147.
- 22. Graesser A.C. et al. AutoTutor: A tutor with dialogue in natural language // Behavior Research Methods, Instruments, & Computers. 2004. T. 36, № 2. P. 180–192.
- 23. He Y., Hui S.C., Quan T.T. Automatic summary assessment for intelligent tutoring systems // Computers and Education. 2009. T. 53, № 3. P. 890–899.
- 24. Heilman M., Madnani N. ETS: Discriminative edit models for paraphrase scoring // Proceedings of the First Joint Conference on Lexical and Computational Semantics 2012. T. 2. P. 529–535.

- 25. Heilman M., Madnani N. ETS: Domain adaptation and stacking for short answer scoring // Second Joint Conference on Lexical and Computational Semantics. 2013. T. 2. P. 275–279.
- 26. Higgins D. et al. Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring // arXiv preprint arXiv:1403.0801. 2014.
- 27. Hou W.-J., Tsao J.-H. Automatic assessment of students' free-text answers with different levels // International Journal on Artificial Intelligence Tools. $-2010.-T.\ 20,\ No.\ 2.-P.\ 327-347.$
- 28. Kanejiya D., Kumar A., Prasad S. Automatic evaluation of students' answers using syntactically enhanced LSA // Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing. -2003. T. 2. P. 53–60.
- 29. Landauer T.K. et al. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans // Proceedings of the 19th annual meeting of the Cognitive Science Society. 1997. P. 412–417.
- 30. Leacock C. Scoring free-responses automatically: A case study of a large-scale assessment // Examens. -2004. T. 3, N_2 1. P. 1–9.
- 31. Leacock C., Chodorow M. C-rater: Automated scoring of short-answer questions // Computers and the Humanities. 2003.-T. 37, N 4. P. 389–405.
- 32. Mitchell T. et al. Towards robust computerised marking of free-text responses // Proceedings of the Sixth International Computer Assisted Assessment Conference. 2002. P. 233–249.
- 33. Mitchell T. et al. Computer based testing of medical knowledge // The 7th computer assisted assessment conference. 2003. P. 249–267.
- 34. Mohler M., Bunescu R., Mihalcea R. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. P. 752–762.
- 35. Mohler M., Mihalcea R. Text-to-text semantic similarity for automatic short answer grading // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. 2009. P. 567–575.
- 36. Pérez D. et al. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment // Revista signos: estudios de lingüística. − 2005. − T. 38, № 59. − P. 325–343.
- 37. Pérez D., Alfonseca E. Adapting the automatic assessment of free-text answers to the students // Proceedings of the 9th CAA Conference, Loughborough: Loughborough University. 2005.
- 38. Pérez D., Alfonseca E., Rodríguez P. Application of the BLEU method for evaluating free-text answers in an e-learning environment // LREC. -2004. -P. 1351-1354.
- 39. Pulman S.G., Sukkarieh J.Z. Automatic short answer marking // Proceedings of the second workshop on Building Educational Applications Using NLP. -2005. -P. 9–16.
- 40. Rosé C.C.P. et al. A hybrid text classification approach for analysis of student essays // Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing. 2003. T. 2. P. 68–75.
- 41. Selvi P., Bnerjee A.K. Automatic short -answer grading system (ASAGS) // arXiv preprint arXiv:1011.1742. 2010.
- 42. Siddiqi R. Improving learning and teaching through automated short answer marking: PhD thesis. University of Manchester, 2010.
- 43. Siddiqi R. Impact of automated short-answer marking on students' learning: IndusMarker, a case study // 2013 5th International Conference on Information and Communication Technologies. 2013.
- 44. Siddiqi R., Harrison C.J., Siddiqi R. Improving teaching and learning through automated short-answer marking // IEEE Transactions on Learning Technologies. 2010. T. 3, N₂ 3. P. 237–249.
- 45. Sukkarieh J., Blackmore J. C-rater: Automatic content scoring for short constructed responses. // FLAIRS Conference. 2009. P. 290–295.
- 46. Sukkarieh J.Z., Pulman S.G., Raikes N. Auto-marking: using computational linguistics to score short, free text responses // The annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK. 2003.
- 47. Sukkarieh J.Z., Stoyanchev S. Automating model building in C-rater // Proceedings of the 2009 Workshop on Applied Textual Inference. Association for Computational Linguistics. 2009. P. 61–69.