УДК 378.168:004

ПРОБЛЕМЫ, ВОЗНИКАЮЩИЕ В ИНТЕЛЛЕКТУАЛЬНЫХ ОБУЧАЮЩИХ СИСТЕМАХ ПРИ ОЦЕНКЕ ОТВЕТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Мишунин О.Б.[1], Савинов А.П.[1], Фирстов Д.И.[1]

[1] *ФГАОУ ВО Национальный исследовательский Томский политехнический университет, Томск, Россия, e-mail: savinov@tpu.ru*

Исследуются возможности системы автоматической оценки свободных ответов на естественном языке, работающей по модели «мешок слов» («bag of words») с использованием Википедии и Рутеза в качестве лингвистических баз знаний. Показано, что этих ресурсов недостаточно для качественной оценки знаний, и необходимо дополнительно разрабатывать онтологию учебного курса. Выявлена необходимость учёта не только состава ключевых слов в тексте, но и грамматических связей между ними. Экспериментально выявлены проблемы, связанные с дублированием или недостаточностью информации в эталоне, что приводит к существенному занижению оценки студенческих ответов. Показано, что возникновения данной ситуации можно избежать следующим образом: 1) эталонный ответ должен содержать полную релевантную вопросу информацию без дублирования; 2) преподавателю следует разбивать эталон на совокупность компонентов, каждый из которых выражает одну законченную мысль и имеет собственную оценку важности; 3) необходимо автоматически определять значимость ключевых слов в компонентах эталона и в ответе студента; 4) в случае неполного ответа система должна генерировать дополнительные вопросы по компонентам эталона, которые не были освещены в ответе студента. В этом случае конечная оценка формируется на основе всех ответов студента на первоначальный и дополнительные вопросы.

Ключевые слова: автоматическая оценка ответов на естественном языке, мешок слов, семантическая близость, автоматическое выделение терминов, интеллектуальные обучающие системы.

PROBLEMS OF AUTOMATIC FREE-TEXT ANSWER GRADING IN INTELLIGENT TUTORING SYSTEMS

Mishunin O. B.[1], Savinov A. P.[1], Firstov D. I.[1]

[1] *National Research Tomsk Polytechnic University, Tomsk, Russia, e-mail: savinov@tpu.ru*

In this paper, we explore the capabilities of the free-text answer grading system based on “bag of words” model with Wikipedia and Ruthes as linguistic knowledge bases. We show that these resources alone are not sufficient for efficient knowledge assessment and development of additional course ontology is needed. It was found necessary to extract grammatical relations between key words instead of using pure “bag of words” model. Experimentally we found problems connected with redundancy or lack of information in a model answer that lead to significant grade drop. Our suggestions on solving these problems are: 1) reference answer should be complete and only contain relevant information; 2) teacher should split the reference answer into a set of components each representing a complete statement and having its own estimated weight; 3) weight of words in student and reference answers should be estimated by system; 4) in case of incomplete student answer the system should be able to generate additional questions on reference answer components skipped by student. In this case, final grade is based on all of the student answers to main and additional questions.

Key words: automatic free-text answer scoring, bag of words, semantic similarity, automatic terms extraction, intelligent tutoring systems.

Работы в области компьютерного обучения, проводимые в Томском политехническом университете, направлены на разработку интеллектуального модуля для LMS Moodle, основной задачей которого является автоматический контроль знаний студентов путём оценки свободных ответов на естественном языке — так, как это происходит при индивидуальном обучении с преподавателем.

Исходной концепцией вышеупомянутых исследований послужили результаты, полученные в работе [2], из которой следовало, что для вопросно-ответных систем

характерно наличие единого терминологического словаря, используемого как для формулирования вопроса, так и для генерации ответа. Возможность перенесения указанных результатов на диалоговую систему контроля знаний изложена в работах [6, 5]. В частности, в работе [5] было указано, что для реализации такой системы необходимо проделать предварительную работу по преобразованию текстов к формализованному виду.

Работы по созданию программных средств, направленных на преобразование текстов на естественном языке в унифицированный формализованный вид, инвариантный 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 нахождения семантической близости отдельных терминов, входящих в текст эталонного ответа и ответа студента. Под терминами в данном случае понимаются как отдельные слова, так и многословные именные конструкции, описывающие одно понятие предметной области.

В этой статье мы исследуем возможность применения модели «мешок слов» с использованием Википедии и тезауруса Рутез для поиска терминов в эталоне и ответе студента и определения их семантической близости. Как показано в обзорах [9, 10], Википедия и тезаурусы (например, WordNet), используются для решения задачи оценки свободных ответов, однако для русского языка подобных исследований не проводилось.

Исходные данные

В качестве предметной области, на которой проводились исследования, был выбран курс «Экономика предприятия» [1]. Этот курс привлёк наше внимание по следующим причинам:

1. Данный курс читается практически для всех специальностей ТПУ, поэтому у нас была возможность получить ответы большого количества студентов для статистического анализа работы системы автоматической оценки.
2. Лекционный материал достаточно хорошо формализован и структурирован. Даются чёткие определения понятиям и терминам, которые имеют похожие формулировки в разных учебниках.
3. С одной стороны, в изложении преобладает текстовый материал, так что системе оценки ответов не нужно работать с математическими формулами. С другой стороны, язык изложения не изобилует сложными оборотами, как, например, в учебниках истории, что снижает вероятность ошибок автоматического анализа естественно-языковых текстов.
4. В LMS Moodle ТПУ создан электронный учебный курс по данному предмету, который можно использовать для апробации результатов нашей работы.

Ниже приведены наиболее интересные примеры вопросов и эталонных ответов из разделов «Основные фонды» и «Оборотные фонды» этого курса. На этих примерах наиболее

ярко демонстрируются проблемы, возникающие при работе метода автоматической оценки свободных ответов.

A. *Перечислите обобщающие показатели использования основных фондов.*

Эталонный ответ: *Фондоотдача, фондоёмкость, фондовооружённость.*

B. *Дайте определение термину «амортизация».*

Эталонный ответ: *Амортизация — это планомерный процесс переноса стоимости средств труда по мере их износа на производимый с их помощью продукт.*

C. *Дайте определение термину «фонды обращения».*

Эталонный ответ: *Фонды обращения — это оборотные средства, обслуживающие процесс реализации готовой продукции; служат для обеспечения непрерывности процесса производства и реализации продукции предприятия (примеры: готовая продукция на складе, товары, отгруженные заказчикам, но ещё не оплаченные ими, дебиторская задолженность, средства в расчётах, денежные средства в кассе предприятия и на счетах в банках).*

D. *По каким объектам основных фондов амортизация не начисляется?*

Эталонный ответ: *Амортизация не начисляется по следующим объектам основных средств: объектам, полученным по договору дарения и безвозмездно в процессе приватизации; жилищному фонду (кроме объектов, используемых для извлечения дохода); объектам, потребительские свойства которых с течением времени не изменяются.*

Алгоритм оценки ответов

Алгоритм оценки ответов системой контроля знаний на естественном языке основан на модели представлении текста как неупорядоченного набора ключевых слов без учёта грамматики и порядка слов, называемой в литературе моделью «мешок слов» («bag of words»). Основными операциями алгоритма являются выделение терминов и оценка их семантической близости. Для этих целей на основе Википедии [4] и Рутеза [3] были сделаны две графовых базы данных.

В узлах графа, построенного по Википедии, находились заголовки статей. Рёбра — ссылки между статьями с указанием их типа: обычная ссылка, ссылка «См. также», ссылка на категорию и т. д. При этом термином считается любой текст, полностью находящийся в узле графа (то есть заголовок одной из статей Википедии целиком). Заголовок страницы-перенаправления считался синонимом термина, на который происходит перенаправление. Семантическая близость между остальными терминами рассчитывалась по формуле Дайса:

$$r = \frac{2|N(a) \cap N(b)|}{|N(a)| + |N(b)|},$$

где $N(a)$ — множество статей, на которые ссылается статья про термин a или в которых есть ссылки на статью про термин a. При этом использовались весовые коэффициенты ссылок, предложенные Д. Ю. Турдаковым [7].

В узлах графа, построенного по Рутезу, находились все текстовые входы тезауруса, а в качестве рёбер использовались связи «выше — ниже». Семантическая близость по Рутезу определялась в соответствии со следующим алгоритмом:

1. Рассчитывается глубина (расстояние от корня) ближайшего предка обоих терминов.
2. Если общий предок найден, семантическая близость вычисляется по формуле:

$$r = \frac{2p_c}{p_1 + p_2},$$

где p_c — путь от корня до общего предка, p_1 — путь от корня до первого термина, p_2 — путь от корня до второго термина.

3. Если у понятий нет общего предка, близость равна 0.

Оценка ответов осуществлялась по следующему алгоритму:

1. Проводится анализ текста эталонного ответа и ответа студента с помощью семантико-синтаксического анализатора ABBYY Compreno [8]. Результат анализа — синтаксическое дерево.
2. Выделяются слова и словосочетания и ставятся в начальную форму.
3. Список слов и словосочетаний фильтруется с помощью Википедии и Рутеза — остаются лишь те слова и словосочетания, которые есть в соответствующих базах.
4. Рассчитывается значение семантической близости между терминами эталона и ответа студента. Строится матрица, в которой строкам соответствуют термины эталона, столбцам — термины студенческого ответа, а в ячейки записывается рассчитанное значение близости. Затем значения в каждой строке сортируются по убыванию близости, а сами строки сортируются по убыванию первого значения в строке. Далее последовательно для каждого эталонного термина выбирается наиболее близкий термин из ответа студента (он будет стоять первым в отсортированной строке). Один и тот же термин студента не может быть выбран дважды, то есть если этот термин из ответа студента уже оказался ближайшим к какому-то термину эталона, то будет взят следующий студенческий термин в строке и т. д. Таким образом, по эталонному ответу и ответу студента получаем список пар наиболее близких по смыслу терминов. Пары, в которых близость терминов была ниже 0,75, отбрасывались, так как было экспериментально установлено, что такие термины обычно не являются близкими по смыслу.
5. Конечная оценка ответа вычисляется по формуле:

$$M = \frac{\sum_{i=1}^{n} r(t_i, s_i)}{|T|},$$

где T — это множество терминов, выделенных из эталонного ответа, t_i — i-й элемент множества T, s_i — термин, выделенный из ответа студента, стоящий в паре с термином t_i, $r(t_i, s_i)$ — значение семантической близости между двумя терминами.

Условия проведения экспериментов

Статистический анализ качества работы интеллектуальной системы контроля знаний обучаемого на естественном языке производился на четырёх изложенных выше вопросах с выборкой ответов на каждый из них, данных 112 студентами. Использовалась только дуальная оценка: «правильно» и «неправильно». Машинные оценки ответов студентов сравнивались с оценками преподавателей, выставленных им как обычно — на основании анализа смыслового содержания текстового ответа обучаемого.

Качество автоматической оценки определялось следующим образом. Всё множество ответов A было разбито на четыре подмножества:

1. Множество ответов, которые были оценены преподавателем как правильные ($A_{C,T}$).
2. Множество ответов, которые были оценены преподавателем как неправильные ($A_{W,T}$).
3. Множество ответов, которые были оценены машиной как правильные ($A_{C,M}$).
4. Множество ответов, которые были оценены машиной как неправильные ($A_{W,M}$).

По этим множествам вычислялись три меры. Первая мера — процент совпадения ответов, оценённых и машиной, и преподавателем как правильные:

$$\epsilon_C = \frac{\left|A_{C,T} \cap A_{C,M}\right|}{\left|A_{C,T}\right|} \cdot 100\%,$$

Вторая мера — процент совпадения ответов, оценённых и машиной, и преподавателем как неправильные:

$$\epsilon_W = \frac{\left|A_{W,T} \cap A_{W,M}\right|}{\left|A_{W,T}\right|} \cdot 100\%,$$

Третья мера — общий процент совпадения оценок машины и преподавателя:

$$\epsilon = \frac{\left|A_{C,T} \cap A_{C,M}\right| + \left|A_{W,T} \cap A_{W,M}\right|}{|A|} \cdot 100\%,$$

Результаты обработки ответов студентов сводились в таблицу, пример которой представлен на Рис. 1. Преподаватели ставили оценку 1 («правильно») или 0 («неправильно»). Машина выставляла оценку от 0 до 1: меньше 0,5 — «неправильно», 0,5 и больше — «правильно». Зелёным подсвечивались строки, в которых оценки преподавателя и машины совпали, красным — несовпадения. Фиолетовым подсвечивались те оценки, которые различались у разных преподавателей. Жирным начертанием в ответах отмечались

однословные термины. Составные термины помечались жирным курсивом. При нажатии на термин можно увидеть все найденные его синонимы (Рис. 2). Тут следует пояснить, что мы намеренно не разрешали лексическую омонимию, так как маловероятно, что студенты и преподаватели употребят один и тот же термин в разных значениях. При нажатии на кнопку «Расчёт» выдавалось окно с рассчитанными коэффициентами семантической близости между найденными в эталонном ответе и ответе студента терминами (Рис. 3).

	ID студента	Оценка преп.	Оценка сист.	Близость
Вопрос: К обобщающим показателям использования основных фондов относятся				
Эталон: **фондоотдача, фондо емкость,** фондовооруженность.				
Процент совпадения: 98,1				
Процент совпадения верных ответов: 100				
Процент совпадения неверных ответов: 96,5				
фондо емкость, фондоотдача, фондовооруженность	118	1	1	Расчёт
фондо емкость	130	0	0,67	Расчёт
***коэффициент износа,* коэффициент оборота, фондо ёмкость,** фондовооруженность.	148	1	0,67	Расчёт

Рис. 1. Пример результатов обработки

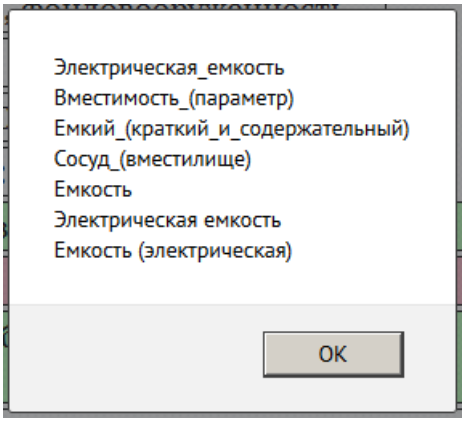

Рис. 2. Пример выделения терминов

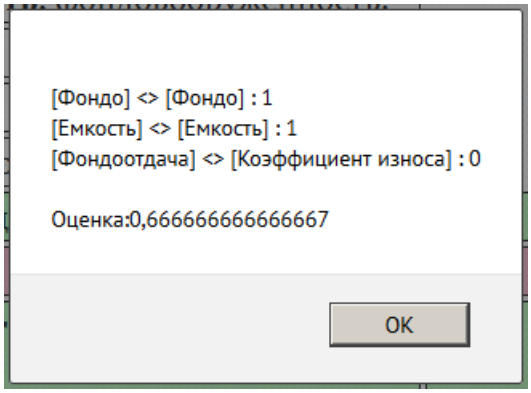

Рис. 3. Пример расчёта семантической близости

Анализ результатов

Вопрос А, $\epsilon = 98{,}1\%, \epsilon_C = 100\%, \epsilon_W = 96{,}6\%$

Ответы, которые требуется дать на вопрос А, по своей конструкции очень похожи на те, которые могут проверяться заложенными в LMS Moodle средствами, например, типом «Краткий ответ». В отличие от стандартного варианта в нашем случае количество ключевых слов в ответе ничем не ограничивается, и преподавателю не требуется записывать различные варианты представления эталона, существование которых обусловлено свободным порядком терминов.

Процент корреляции машинной оценки ответов студентов с оценками преподавателей в данном случае достиг максимальной по сравнению с другими результатами величины. Анализ показывает, что несмотря на наблюдаемую высокую корреляцию, в оценках машины присутствуют искажения, обусловленные неправильным определением терминов в эталоне и ответе студента с помощью Википедии и Рутеза. Из трёх терминов ими был правильно выделен только термин *фондоотдача*. Термин *фондоёмкость* был заменён данными системами на два термина *фондо* и *ёмкость*, а термин *фондовооружённость* в базах Википедии и Рутеза вообще отсутствует. Это привело к искажению расчётов и завышению оценок. Ярким примером является случай, когда в ответе студента был приведён единственный из трёх терминов — *фондоёмкость*. Оценка преподавателя была «0», а система поставила «1», так как интегральный коэффициент семантической близости между терминами, благодаря их раздвоению, составил 0,666. Вывод: использование баз Википедии и Рутеза недостаточно для выделения ключевых слов и вычисления их семантической близости. Необходимо дополнительно разрабатывать онтологию предметной области, в данном случае онтологию курса «Экономика предприятия», с помощью которой можно будет выделять термины и рассчитывать семантическую близость между ними.

Вопрос В, $\epsilon = 84{,}5\%, \epsilon_C = 81{,}0\%, \epsilon_W = 89{,}7\%$

Вопросы типа В — это очень распространенная конструкция вопроса, при ответе на который студенту требуется дать определение какому-либо понятию. В нашем примере — дать определение понятию «амортизация».

Процент совпадения машинной оценки с оценками, поставленными преподавателем, был достаточно высоким. При этом в 8 ответах машинная оценка была завышена (система поставила «правильно» ответу, оценённому преподавателем как неправильный), а в 1 занижена (правильный ответ оценён как неправильный). Это неплохой результат. Следует отметить, что студенты не были предупреждены о том, что их ответы будет проверять компьютер, поэтому отвечали так, как отвечали бы преподавателю. В связи с этим имеющиеся в выборке и приводимые далее абсурдные ответы следует связывать не с попытками студентов обмануть машину, а именно с недостаточным знанием предмета.

Чтобы оценить данную ситуацию, рассмотрим один из вариантов, входящих в состав восьми неправильно оценённых машиной ответов студентов:

Амортизация — планомерный процесс переноса стоимости продукции объектам основных фондов.

Смысл, вложенный студентом в данное предложение, полностью противоречит смыслу, заложенному в эталоне. Система оценки не реагирует на порядок следования ключевых слов в ответе, на оценку влияет только их количество. Данное обстоятельство указывает на

необходимость доработки алгоритма работы системы так, чтобы он учитывал не только состав ключевых слов в тексте, но и взаимную связь между ними.

Вопрос С, $\epsilon = 54{,}1\%, \epsilon_C = 0\%, \epsilon_W = 100\%$

Третий тип задания заинтересовал нас тем, что преподаватель, чтобы обеспечить определённую гибкость эталонному ответу, вложил в него избыточную информацию. Как видно из статистических данных, система сильно занижает оценку. Возникновение данной ситуации, в которой ни один ответ студента не оценен, как правильный, обусловлено тремя причинами:

1. Дублированием информации в эталонном ответе (приведено определение и примеры), что привело к двойному увеличению в нём объёма ключевых слов, на который производится нормирование интегрального коэффициента семантической близости ключевых слов, содержащихся в тексте ответа и в эталоне.
2. Стремлением студентов кратко излагать свои мысли в ответах на поставленный вопрос — студенты обычно приводили что-то одно: или определение, или примеры.
3. Лояльным отношением преподавателей, которых, как правило, устраивает краткий ответ студентов.

В результате этого значение нормированного интегрального коэффициента семантической близости ключевых терминов стало меньше установленного порога равного 0,5 и, соответственно, система существенно занизила оценки ответам студентов по сравнению с преподавателем.

В качестве эксперимента исходный эталон был разделён на две равноправные части:

1. *Оборотные средства, обслуживающие процесс реализации готовой продукции; служат для обеспечения непрерывности процесса производства и реализации продукции предприятия.*
2. *Готовая продукция на складе, товары, отгруженные заказчикам, но ещё не оплаченные ими, дебиторская задолженность, средства в расчётах, денежные средства в кассе предприятия и на счетах в банках.*

В этом эксперименте оценка ответов студентов производилась путём поочередного сравнения их с каждой составляющей разделённого эталона. За конечный результат принимался рассчитанный вариант с максимальным значением интегрального коэффициента семантической близости ключевых терминов. В результате статистического анализа были получены следующие результаты: $\epsilon = 86{,}0\%, \epsilon_C = 55\%, \epsilon_W = 100\%$. Как видно, применённая процедура разделения эталона, дала положительный эффект в определении правильных ответов, который может быть увеличен ещё больше, если в составе системы кроме базы знаний Википедии и Рутеза будет онтология предметной области «Экономика

предприятий», позволяющая анализировать многокомпонентные специализированные термины.

Вопрос D, $\epsilon = 75{,}4\%, \epsilon_C = 12{,}5\%, \epsilon_W = 100\%$

Интересно рассмотреть и четвёртое типовое задание, поскольку оно развивает дальше важную проблему, поднятую при анализе предыдущего примера, а именно проблему формирования эталона, который определяет качество работы системы. Сравнивая результаты статистического анализа с предыдущими, видим, что они очень близки по величине. Наблюдаемое занижение оценки в данных примерах обусловлено разными причинами. Если в предыдущем варианте в эталоне присутствовало дублирование информации, то в последнем случае в эталон была введена неполная информация о том, в каких случаях амортизация не начисляется. Как показал анализ литературы, полный ответ на этот вопрос должен быть примерно на 75% шире. Студенты в своих ответах отражают эти дополнительные по отношению к эталону сведения, а преподаватели, располагая большим багажом знаний, положительно оценивают эти ответы.

Поскольку контролирующая компьютерная система, в отличие от преподавателя, располагает объёмом информации в пределах эталона, то неудивительно, что оценки, сделанные компьютером, расходятся с преподавательскими. Для исключения случаев возникновения данной ситуации необходимо, чтобы эталон содержал полную информацию, соответствующую предъявленному студенту заданию. Однако в этом случае возникает вопрос, как оценивать знания студентов, которые, как правило, кратко излагают свой ответ по сравнению с эталоном, что приводит, в конечном счёте, к ситуации, аналогичной той, что возникает при дублировании информации в эталоне (пример C). Контролирующая система не оценит положительно ни один ответ студента, как и в предыдущем случае.

Чтобы избежать проблем, выявленных при анализе вопросов C и D, предлагается следующая концепция формирования и использования эталона в интеллектуальной системе контроля знаний:

1. Эталонный ответ должен содержать полную релевантную вопросу информацию без дублирования.
2. Преподаватель разбивает эталон на совокупность компонентов, каждый из которых выражает одну законченную мысль. Каждый компонент оценивается им с точки зрения важности.
3. Система автоматически определяет значимость ключевых слов в компонентах эталона и в ответе студента.
4. В случае неполного ответа студента система генерирует дополнительные вопросы по компонентам эталона, которые не были освещены в ответе.

Заключение

В ходе исследования метода оценки свободных ответов на естественном языке, основанного на модели «мешок слов», были сделаны следующие выводы:

1. Модель «мешок слов» в чистом виде не подходит для оценки ответов на большинство типов вопросов. Для повышения качества оценки требуется выделять и использовать грамматические связи между словами.
2. Терминологии в использованных в качестве лингвистических баз знаний Википедии и Рутезе недостаточно для оценки ответов по специальным дисциплинам, и требуется разработка онтологии предметной области.
3. Экспериментально выявлено влияние вида эталона, его содержания и структуры на качество оценки свободных ответов. Предложена концепция формирования эталонного ответа и использования его в диалоговой системе оценки знаний, что должно устранить выявленные в исследовании проблемы.

Список литературы

1. Воробьёва И. П., Селевич О. С. Экономика предприятия: учебное пособие // Томский политехнический университет. — Томск, изд-во Томского политехнического университета, 2013. — 179 с.
2. Кибрик А. Е., Нариньяни А. С., Ершов А. П. Моделирование языковой деятельности в интеллектуальных системах. — М., «Наука», Глав. ред. физико-математической лит-ры, 1987.
3. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. — М.: Издательство Московского университета, 2011. — 512 с.
4. Русская Википедия // Википедия. [2005–2015]. Дата обновления: 15.10.2015. URL: http://ru.wikipedia.org/?oldid=73917479 (дата обращения: 03.11.2015).
5. Савинов А. П. Интеллектуализация обучающей системы Moodle позволит продлить срок эксплуатации ее в вузах / А. П. Савинов, Т. С. Петровская, Д. И. Фирстов // Высшее образование сегодня. — 2014. — № 9. — С. 15–21.
6. Сулейманов Д. Ш. Двухуровневый лингвистический процессор ответных текстов на естественном языке // Сборник трудов Международной научно-технической конференции OSTIS-2011, Минск. — 2011. — С. 311–322.
7. Турдаков Д. Ю. Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов // Дисс. канд. ф-м. наук. Москва. — 2010. — С. 61–65.

8. ABBYY Compreno // ABBYY. [2015]. URL: http://www.abbyy.ru/isearch/compreno/ (дата обращения: 03.11.2015).
9. Burrows S., Gurevych I., Stein B. The eras and trends of automatic short answer grading // International Journal of Artificial Intelligence in Education. — 2015. — Т. 25. — №. 1. — С. 60–117.
10. Pérez-Marín D., Pascual-Nieto I., Rodríguez P. Computer-assisted assessment of free-text answers // The Knowledge Engineering Review. — 2009. — Т. 24. — №. 04. — С. 353–374.

Рецензенты:

Гришин А.М., д.ф.-м.н., профессор, зав. кафедрой физической и вычислительной механики Национального исследовательского Томского государственного университета, г. Томск;

Тузовский А.Ф., д.т.н., профессор кафедры оптимизации систем управления Института кибернетики Национального исследовательского Томского политехнического университета, г. Томск.