УДК 378.146:004

А.А. Мицель, А.А. Погуда

Универсальный алгоритм проверки естественно-языковых текстов

Приводится универсальный алгоритм, предназначенный для автоматизированного контроля знаний для гуманитарных дисциплин. Определяются основные преимущества и недостатки данного алгоритма по сравнению с другими алгоритмами.
Ключевые слова: контроль знаний, модели и алгоритмы, компьютерные системы контроля знаний.

Введение

Традиционные способы контроля и оценивания знаний дистанционно обучаемых студентов путем тестирования сводятся к предъявлению к тестируемому студенту фиксированного множества тестовых заданий и различных вариантов ответов на каждое из них. Задача обучаемого состоит в выборе одного или нескольких истинных, по его мнению, ответов на каждое тестовое задание. Основу этих способов составляет оценивание истинности предлагаемых вариантов ответов «правильно – неправильно», что требует от организатора тестирования признать абсолютную истинность вариантов ответа и абсолютную ложность остальных вариантов. Органический недостаток подобного подхода состоит в невозможности учитывать при тестировании неполные или не совсем точные ответы обучаемого.

Данная проблема была подробно затронута авторами в работе [1], где приводятся конкретные примеры алгоритмов и предлагается новый универсальный алгоритм для дисциплин, в которых ответы на вопросы даются в словесной форме. В данной статье более подробно описываются предложенный в [1] алгоритм, а также некоторые особенности, которые следует учитывать при корректной оценке ответов на вопросы.

Анализ естественно-языковых текстов

Анализ естественно-языковых текстов представляет собой очень актуальную проблему, особенно в последнее время, ввиду большого роста объемов текстовой информации и сложной структурированности естественно-языковых текстов. Существует множество статических и нестатических подходов к поиску текстовой информации. Статические в основном используются для оценки и вычисления релевантности документа запросу, применяемых в современных поисковых системах, относимых к статическому подходу анализа текстовых данных, что не совсем подходит для решения задачи тестирования. Что касается нестатических подходов, то здесь ситуация несколько другая. Нестатические подходы могут быть применены в диалоговых системах при построении ответов на естественно-языковой вопрос, в системах машинного перевода и других видов анализа информации. В работе [2] автор предлагает следующую классификацию сложности решений:

- оригинальные решения. *Класс A*;
- красивые решения. *Класс B*;
- сложные решения. *Класс C*.

К классу *A* относятся решения, в основу которых положен автоматизированный анализ естественно-языковых текстов. Здесь способ получения данных возлагается на вычислительные мощности ЭВМ.

К классу *B* относится создание словарей и фундаментальных работ, которые и легли в основу современной прикладной лингвистики. Решения этого класса подразумевают долгий и трудоёмкий ручной процесс, требующий немалых временных затрат, но результат будет универсальным для большого спектра входных переменных.

К классу *C* относятся более сложные решения, в которых число операций для реализации поставленной задачи существенно превышает число слов в языке. Здесь же больше задействован ручной труд, нежели производительные мощности ЭВМ. К этому классу относится целый ряд методов анализа – лексический, морфологический, синтаксический.

Задача лексического анализа состоит в разделении текста на слова, разделители; в выделении устойчивых оборотов, не имеющих словоизменительных вариантов; выделении фамилии, имени, отчества; числовых и иных знаковых комплексов, предложений,

абзацев и др. Данный анализ вырабатывает информацию, которая передается на последующие этапы обработки, т.е. морфологическому и синтаксическому анализаторам.

Задача морфологического анализа состоит в однозначном определении леммы (начальной формы слова) и парадигмы (всех грамматических словоформ для леммы) для каждого из слов в анализируемом предложении.

Задача синтаксического анализа состоит в выделении в предложении синтаксических единств (фрагментов), больших или равных словосочетанию (синтаксической группе), и в установлении иерархии этих единств, без использования семантической информации и информации о модели управления. Иерархия здесь отражает синтаксическую зависимость отдельных фрагментов в предложении.

Алгоритм для проверки естественно-языковых текстов

Существует множество алгоритмов для решения задачи автоматизированного контроля проверки знаний. Предлагаемый метод подразумевает комбинированное использование лексического, морфологического и синтаксического анализов. Пусть «Перечислите основные компоненты системного блока современного персонального компьютера» будет нашим вопросом, на который отвечает тестируемый, а «Основными компонентами системного блока современного компьютера являются: материнская плата, процессор, оперативная память, жесткий диск, видеокарта, устройство чтения диска» будет нашим полным ответом. То есть полным ответом служит фраза, которую обычно требуют преподаватели при устном экзамене.

Хочется отметить, что ответами на такие вопросы будут служить основные ключевые слова или словосочетания, следующие по убыванию по степени важности, т.к. в будущем разрабатываемая нами система оценки будет руководствоваться как раз основными моментами, чтобы в полной мере оценить ответ. Как показывает практика с ЕГЭ или распространенной в данный момент среди вузов системой «Интернет-экзамен в сфере профессионального образования», здесь вопросы сконцентрированы на комбинации ключевых слов, среди которых только одна из них правильная. В таких случаях экзаменуемый, даже зная правильный ответ, затрудняется ответить на вопрос. На самом деле, при ответе на данный вопрос порядок ключевых слов в комбинации не должен иметь значения, т.е. любая комбинация должна оцениваться как правильный ответ. А связано это с тем, что обычно ответом на такие вопросы служат фразы, в целом похожие друг на друга, т.е. можно сказать и так, и так, но правильной, как правило, является только одна формулировка. Вот здесь и возникает проблема, потому что при устном экзамене, если экзаменуемый ответит, как он понял данный материал и где-нибудь ошибется, то преподаватель сможет поправить и направить мышление экзаменуемого в нужное русло. В этом случае получается, что ответ все же был правильным, пусть и частично, но правильным. В итоге преподаватель, анализируя данный ему ответ, решает задать дополнительный вопрос или же экзаменуемый справился с поставленным ему вопросом. В том или ином случае преподаватель ставит заслуженную оценку. Но и здесь не все так просто! Данная оценка может зависеть и от сторонних признаков, например душевного или психологического состояния преподавателя. Ни для кого не секрет, что у любого, но не каждого преподавателя есть свои «любимчики», с которыми они стараются или больше работать, или уделять им основное внимание. Таким «любимчикам» в каком-то роде проще сдать экзамен, так как даже если ответ будет дан неполным или неверным, то преподаватель постарается «вытянуть» на хорошую оценку дополнительными вопросами или даже подсказками. Из всего вышесказанного, можно сделать вывод, что не следует «навязывать» шаблоны тестируемому, а это в будущем, возможно, полностью поменяет его мировоззрение.

Итак, возьмем слова в нашем вопросе и ответе за переменные, где слова во всех падежах и временах представляются в виде определенного набора букв: a = «перечислить», b = «основные», c = «компоненты», d = «персональный», e = «компьютер», f = «материнская плата», g = «процессор», h = «оперативная память», i = «жесткий диск», j = «видеокарта», k = «устройство чтения дисков». При построении формулы вопроса и ответа следует учитывать даже самые тонкие моменты, т.к. от них может зависеть дальнейшее развитие событий. Здесь имеется в виду, что мы ввели разными переменными фразу «персональный компьютер», она у нас состоит из формулы $d + e$, хотя могли ввести и одну переменную. А сделали мы это потому, что в последнее время индустрия компьютерных технологий активно развивается, и появилось такое понятие, как «мобильный компьютер», т.е. ноутбук или даже нетбук. В нашем случае, конечно, разница не существенна, но на будущее это стоит учитывать. Однако если взять мобильные компью-

теры, а в частности нетбук, то у них отсутствует устройство чтения дисков, поэтому при полном ответе данное устройство можно не учитывать.

В данном случае ключевым ответом на наш вопрос будет, обязательное, наличие переменных f, g, h, i, j и k в ответе. Предполагается, что используется база данных слов, в которой содержатся все возможные слова, применяемые в ответе в разных падежах и временах, а также, возможно, и их синонимы и даже жаргонный сленг. Так, например, для переменной i и j буде верным такой набор слов: i = (жесткий диск, HDD, винт, винчестер, устройство хранения информации), j = (видеокарта, видеоадаптер, видео). Тогда формула нашего вопроса будет выглядеть так:

$$F = a + b + c + d + e, \quad (1)$$

где F – вопрос, $a + b + c + d + e$ – наши переменные.

Ответами будут являться множество O_1, O_2, O_3, ... ,O_n, где полным ответом будет считаться O_1:

$$O_1 = b + c + d + e + f + g + h + i + j + k. \quad (2)$$

Другими же неполными, но верными, ответами будут являться:

$$O_2 = b + c + d + e + f + g + h + i + j,$$
$$O_3 = b + c + d + e + f + g + h + I,$$
$$\ldots,$$
$$O_n = f.$$

Для выведения формулы алгоритма приведем все переменные нашего вопроса и ответа к единым переменным, тогда вопрос a, b, c, d, e будет выглядеть как a_1, a_2, a_3, a_4, a_5, а ответ f, g, h, i, j, k будет выглядеть как b_1, b_2, b_3, b_4, b_5, b_6. В результате получим:

$$F = a_1 + a_2 + a_3 + a_4 + a_5,$$
$$O_1 = a_2 + a_3 + a_4 + a_5 + b_1 + b_2 + b_3 + b_4 + b_5 + b_6,$$
$$O = \sum_{i=1}^{n} a_i + \sum_{j=1}^{m} b_j . \quad (3)$$

Очевидно, что также требуется учитывать перестановку переменных b_j. В нашем случае как раз приведен подобный пример, когда требуется реализация условия перестановки, т.к. в нашем случае, перестановка переменных b не зависит от правильности ответа. Но так как этот параметр подходит не для всех видов вопросов, например требуется перечислить иерархию правления на Руси, то его требуется сделать отключаемым в каждом конкретном случае.

В результате если тестируемый ответит на вопрос правильно и полным ответом, система засчитает ответ правильным и дополнительно выставит оценку ответа в процентном соотношении, т.е. 100%. В случае если ответ будет верным, но не полным, например «Основными компонентами системного блока современного компьютера являются: материнская плата, процессор, оперативная память, жесткий диск, видеокарта», т.е. вариант ответа O_2, тогда он будет оценен в 90%, аналогично, если экзаменуемый ответит «Материнская плата, процессор, оперативная память, жесткий диск, видеокарта», ответ будет также оценен в 90%, т.к. он будет содержать основную формулировку ответа. Также если ответ будет звучать «Процессор, жесткий диск, оперативная память, материнская плата», ответ будет оценен на 80% и т.д. Стоит помнить, что система будет основываться на ответе, который дал тестируемый в процентном соотношении, поэтому для каждой дисциплины преподаватели могут сформировать шкалу ответов, где, например, ответом на вопрос, данный в 90–100%, будет засчитываться оценка «отлично», в 60–80% – «хорошо», 30–50% – «удовлетворительно». Если же используется комплекс вопросов, то итоговую оценку можно выводить как среднестатистическую из всех вопросов, на которые ответил экзаменуемый, и ее считать конечным результатом.

Если же ответ дан неверно, то система ответ не засчитает, но если в ответе встречаются ключевые слова, то оценка ответа в процентном соотношении будет равна количеству встречающихся ключевых слов, но без повторений. Данный результат зафиксируется в системе и будет выведен при просмотре результатов в отдельную таблицу, где будут содержаться ответ, данный экзаменуемым, и его оценка.

В том или ином случае все ответы и оценки по каждому тестируемому будут формироваться в таблицы в отдельные шифрованные файлы, которые смогут просмотреть только администраторы. Данная функция необходима в случае возникновения спорных ситуаций.

Также требуется разработать систему проверки ошибок ввода, это можно реализовать 2 способами:

– ввести систему проверки ввода в поле ввода ответа;

– все ответы, которые были даны с ошибками, заносить в ту же таблицу, в которой выводятся неверные ответы.

Таким образом, возможен еще один метод тестирования, преподаватель не только сможет следить за успеваемостью, но и видеть, по какой именно теме у обучающихся имеются «пробелы», а также учить тестируемых студентов правильно формировать ответ.

Заключение

Предложен новый алгоритм для проверки знаний учащихся по гуманитарным дисциплинам. Как и в любом другом алгоритме [1], можно выделить преимущество и недостатки. Достоинство предлагаемого алгоритма: универсальность при обработке естественно-языковых текстов. Недостаток: данный алгоритм не сможет обрабатывать тригонометрические формулы и функции, использующиеся в высшей математике, математической статистике и эконометрике.

Литература

1. Мицель А.А. Тестирование как неотъемлемая часть современного образования / А.А. Мицель, А.А. Погуда // Матер. докл. VI Всерос. науч.-практ. конф. студ., асп. и молодых ученых «Инноватика-2010». Томск, 12–17 апреля 2010 г. – Томск: ТГУ, 2010. – Т. 2. – С. 92–96.

2. Борисов А.Н. Принятие решений на основе нечетких моделей / А.Н. Борисов, О.А. Крумберг, И.П. Федоров. – Рига: Знание, 1990. – 352 с.

Мицель Артур Александрович
Д-р техн. наук, проф. каф. автоматизированных систем управления (АСУ) ТУСУРа
Тел.: (382-2) 70-15-36
Эл. почта: maa@asu.tusur.ru

Погуда Алексей Андреевич
Аспирант каф. АСУ ТУСУРа
Тел.: (382-2) 52-95-35
Эл. почта: alexsmail@sibmail.com

Mitsel A.A., Poguda A.A.

Universal check algorithm for natural language texts

The manuscript presents the universal algorithm to check knowledge automatically in courses, where strong dialectical method is required, and defines its principle advantages and disadvantages in comparison with other algorithms.

Keywords: the knowledge control, models and algorithms, computer systems of knowledge control.
