

УДК 004.896:519.767.6

А.М. Куртасов, А.Н. Швецов

МЕТОД АВТОМАТИЗИРОВАННОЙ ГЕНЕРАЦИИ КОНТРОЛЬНО-ТЕСТОВЫХ ЗАДАНИЙ ИЗ ТЕКСТА УЧЕБНЫХ МАТЕРИАЛОВ

Предложен метод генерации контрольно-тестовых заданий путем обработки естественно-языковых текстов. Описаны алгоритмы и архитектура генерирующей системы, использующей модель оценки качества получаемых заданий на основе машинного обучения.

Тесты проверки знаний, генерация тестовых заданий, генерация вопросов, обработка естественного языка, машинное обучение.

The method for generating the control and test items by processing natural-language texts is proposed in the article. The paper describes the algorithms and architecture of the generating system using the learning-based model for evaluating the quality of the produced test items.

Knowledge assessment tests, test item generation, question generation, natural language processing, machine learning.

Введение. Состояние проблемы. Контроль знаний является фундаментальной функцией образования, реализация которой требует значительных трудозатрат от преподавателей. Системы электронного обучения облегчают задачу контроля знаний за счет средств автоматизированной проверки знаний в форме тестирования. Такие средства обеспечивают проведение тестов в автоматическом режиме, в том числе дистанционно, однако процедуры композиции тестов до сих пор автоматизированы слабо: содержимое в основном создается вручную. Подготовка контрольно-тестовых заданий (КТЗ) представляет собой сложный процесс анализа, подготовки и обработки информации.

Анализ публикаций в области автоматизированной подготовки тестов [5, с. 73–118], [6] позволяет выделить три основных направления исследований. Значительное число научных работ связано с применением текстовых шаблонов КТЗ и их обработкой при помощи различных математических аппаратов (формальные грамматики, деревья И/ИЛИ и т. п.). Такие методы обеспечивают высокую вариативность тестов, однако составление шаблонов, как правило, осуществляется вручную.

Отдельные работы посвящены генерации КТЗ на основе инженерии знаний: предметная область (ПО) представляется в виде некоторой модели (например, семантической сети), из которой затем строятся КТЗ. В системах такого рода возможно достаточно точное соответствие между порождаемыми КТЗ и концептуальными сущностями, составляющими ПО, однако построение моделей ПО представляет собой отдельную сложную задачу.

Третий подход к генерации КТЗ, исследуемый в настоящей работе, основан на извлечении текстовых фрагментов из электронных документов (учебников,

статей и т.п.). Идея подхода заключается в том, что за счет средств обработки естественного языка (ОЕЯ) из составляющих текст предложений можно в автоматизированном режиме генерировать вопросы или упражнения того или иного вида.

Исходная концепция подхода применительно к русскому языку изложена в работе [3] на примере текстов по аэрокосмической тематике. В зарубежных работах данная тема освещена шире: например, описаны исследования, посвященные таким аспектам проблемы, как: оценка качества генерируемых из текста вопросов [8] и генерация ответов-дистракторов для КТЗ закрытой формы при помощи системы *WordNet* [10]. Авторы публикаций характеризуют проблему генерации вопросов как нетривиальную и указывают на необходимость учета семантики при обработке естественно-языковых текстов.

В работах [7], [9] дано описание базовых алгоритмов генерации КТЗ, используемых в разрабатываемом методе, и обозначены проблемы, связанные с семантической обработкой текста. Наиболее существенными представляются следующие проблемы: а) выделение самых важных с точки зрения смысла предложений из текста и б) выделение в предложении слов и фраз, к которым следует обращать вопросы (далее будем называть такие слова и фразы *ключами*).

Перспективным подходом к решению указанных проблем может являться машинное обучение (МО) – ряд статистических методов, позволяющих строить гипотезы о зависимостях между переменными, явные связи для которых не установлены. Пригодность предложений в проблеме (а) и ключей в проблеме (б) будет скрытой переменной, а набор наблюдаемых переменных (т.е. пространство признаков) будет складываться из информации об элементах текста,

которую можно получить доступными средствами ОЕЯ при обработке текста.

Схема реализации метода. Практическим воплощением метода должна стать генерирующая система, получающая на входе произвольные естественно-языковые (ЕЯ) тексты и предлагающая пользователю варианты КТЗ по данному тексту. Архитектура системы представлена на рис. 1. Модули системы решают следующие задачи:

- 1) обработка текста для получения его формализованного представления;
- 2) генерация КТЗ при помощи алгоритмов и правил трансформации;
- 3) классификация получаемых КТЗ по пригодности на основе предшествующего опыта и информации, получаемой на этапе (1);
- 4) фильтрация набора КТЗ с удалением непригодных КТЗ по результатам этапа (3).

Формализация задачи. Тест проверки знаний представляет собой некоторое упорядоченное множество КТЗ. В составе тестирующих систем обычно присутствует банк КТЗ – неупорядоченная совокупность всех КТЗ, из которых могут формироваться тесты. Представим банк заданий как множество T тестовых заданий t :

$$T = \{t_i : 1 \leq i \leq m\},$$

где m – число заданий в банке, или мощность банка заданий. Тогда формируемые тесты T_k будут представлять собой упорядоченные наборы вида:

$$T_k = (t_1, \dots, t_n), \quad 1 \leq n \leq m, \quad \forall (i \neq j), \quad t_i \neq t_j,$$

где n – число заданий в тесте.

Для формирования тестов из банка заданий требуется реализация инъекции $(1, \dots, n) \rightarrow T$. На состав тестов могут накладываться ограничения, исходя из атрибутов, приписанных отдельным КТЗ (например, относительная сложность, принадлежность к той или иной теме и т.п.). Каждое КТЗ представим в виде следующего кортежа:

$$t = (q, s, k, A, P),$$

где q – основа задания, s – источник задания, k – ключ, A – множество вариантов ответа, P – множество педагогических атрибутов. Под основой задания q мы понимаем текст, предлагаемый испытуемому (вопросительная конструкция или задание на заполнение пропусков). Член s вводится для сохранения исходной фразы, из которой порождено задание. Ключ k – слово или фраза, к которой задается вопрос в данном КТЗ (в случае открытого задания соответствует строке правильного ответа, а в случае закрытого задания сопоставляется одному из вариантов ответа). Множество вариантов ответа $A = \{k\} \cup D$, где D – множество дистракторов, которое содержит отвлекающие варианты ответов в случае закрытых заданий; $D \neq \emptyset$ в случае открытого задания. Множество P содержит атрибуты, получаемые после расчета тестологических показателей задания (например, относительная сложность), а также приписываемые ему в рамках данного учебного курса (принадлежность теме); непосредственно после генерации КТЗ $P = \emptyset$.

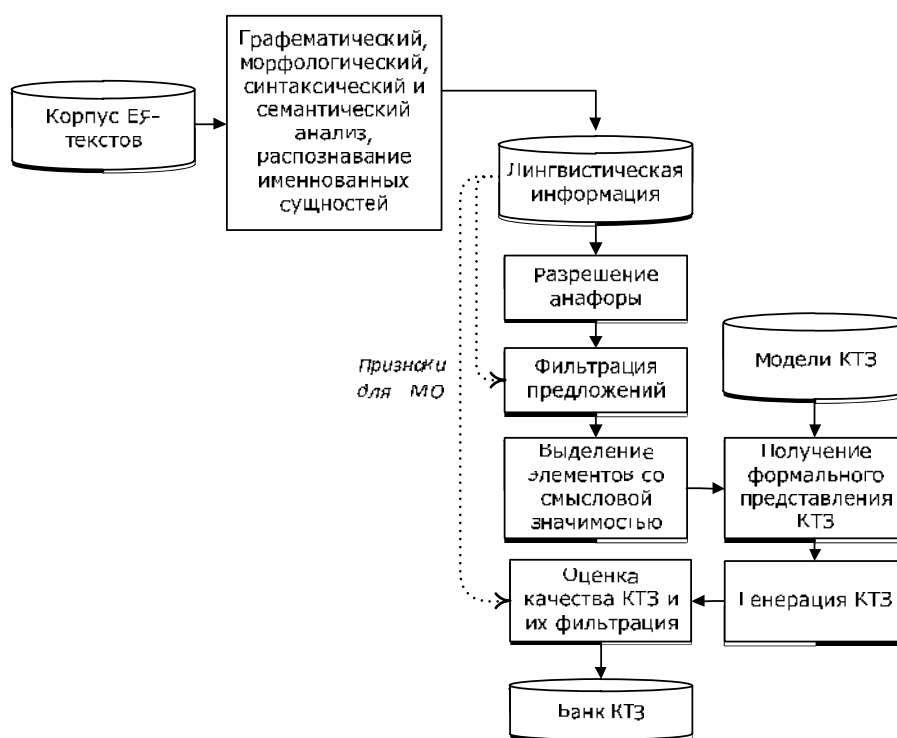


Рис. 1. Архитектура генерирующей системы

Следуя введенным обозначениям, процедуру генерации КТЗ абстрактно представим как преобразование

$$G : C \rightarrow (t_1, \dots, t_p),$$

а процедуру фильтрации КТЗ как

$$F : (t_1, \dots, t_p), X \rightarrow (t_1, \dots, t_q), q \leq p,$$

где C – обрабатываемый корпус текстов, p – число КТЗ, порожденных после выполнения процедуры генерации, q – число КТЗ, прошедших фильтрацию, X – матрица значений признаков, по которым классифицирующий алгоритм осуществляет фильтрацию.

Преобразование F осуществляется на базе вычисления значения функции h – гипотезы классификации, получаемой при помощи машинного обучения из предшествующего опыта работы системы (рис. 2) [11]. Матрица X формируется из данных, получаемых от средств ОЕЯ при анализе корпуса C . Размерность матрицы $p \times f$, где f – число признаков, присущих каждому сгенерированному КТЗ.



Рис. 2. Определение пригодности КТЗ на основе машинного обучения

Обработка текста. Цель обработки текста – преобразование текста в лексемно-морфологические, синтаксические и семантические представления, позволяющие осуществлять необходимые преобразования текста и получать об элементах текста лингвистическую информацию, используемую в качестве признаков для машинного обучения.

Предварительный этап – *подготовка документов* – заключается в выполнении поверхностных операций над текстом с целью получить структурированный набор текстовых данных. В число этих операций входят идентификация кодировки символов, идентификация языка (в данном случае русского), секционирование текста (определение фактического содержимого документов с удалением ссылок, элементов форматирования и т.п.).

Следующим этапом является *сегментация текста*, или *графематический анализ*. Основными задачами здесь являются токенизация (разбиение текста на орфографические слова, или токены), нормализация (приведение различных форм написания токенов к единой форме) и сегментация на предложения с учетом различных нюансов (например, точка не всегда указывает на конец предложения). Мы принимаем утвердительные предложения текста за основные единицы для формирования КТЗ, так как предложение по определению служит для формирования и выражения законченной мысли в речи. Поэтому этап

сегментации на предложения представляет большую важность для генерации КТЗ.

При *морфологическом анализе* для каждого использованного в тексте слова определяется набор морфологических характеристик: часть речи, род, число и падеж для существительных, число и лицо для глаголов и т.п. Этап сопряжен с рядом проблем, связанных с лексико-морфологической омонимией, которые в современных реализациях решаются за счет эвристических и вероятностных моделей. Применяв морфологический анализатор к обрабатываемому тексту, мы получаем набор морфологических признаков для каждого токена.

Синтаксический анализ имеет целью определить роли слов в предложении и их связи между собой. Проблемой этапа является возникновение альтернативных вариантов синтаксического разбора вследствие несовершенства правил, по которым обнаруживаются связи между словами, и наличия лексико-морфологической многозначности словоформ. Результатом анализа является представление синтаксических связей каждого предложения в виде некоторой модели. Мы применяем модель дерева составляющих, в которой структура предложения представляется как иерархия групп (именных, глагольных, предложных и т.п.) и входящих в группы вершин, представляющих слова с морфологической информацией.

Цель *семантического анализа* – установление смысловых отношений между словами в тексте. В основе анализа лежит идея о том, что значение слова не является элементарной смысловой единицей [1]. Оно делимо на более элементарные смыслы, которые составляют словарь особого семантического языка. В результате анализа получается семантический граф, показывающий отношения между смысловыми единицами. Поскольку такой граф является отображением слов в смыслы, информацию из него целесообразно использовать для выделения содержательно-значущих элементов.

Еще один важный этап обработки текста заключается в *выделении именованных существностей* (персоналий, географических названий, дат и т.п.). Как правило, эти существности представляют смысловую ценность, поэтому их выделение целесообразно использовать в алгоритмах генерации КТЗ.

После получения морфосинтаксической и семантической информации о предложениях выполняются две процедуры обработки текста, направленные на получение набора предложений, готовых к преобразованию в КТЗ: разрешение анафоры и фильтрация предложений.

Разрешение анафоры, или референциальный анализ, необходим потому, что в текстах нередко используются языковые выражения, которые не могут быть проинтерпретированы без обращения к предыдущему контексту, например, местоимения третьего лица. Разрешение анафоры позволяет установить, к какому языковому выражению (антецеденту) из предыдущего текста относится анафорическое местоимение и к какой существности реального мира (рефе-

ренту) местоимение и его антецедент отсылают [2]. В нашей задаче разрешение анафоры целесообразно задействовать для автоматической замены местоимений на соответствующие существительные при извлечении предложений.

Фильтрация предложений является одной из ключевых проблем в задаче генерации КТЗ. В системах ОЕЯ процедура сокращения текста с целью сохранения наиболее важных его частей называется автоматическим реферированием. К задаче выбора полезных с точки зрения смысла предложений можно применить частный случай автореферирования, когда из текста выделяются характерные фрагменты без формирования связанных аннотаций (составление выдержек). В число параметров, рассматриваемых при фильтрации предложений, входит расположение предложения в тексте, вхождение в предложение ключевых фраз (например, «в заключение»), статистические меры частоты появления определенных терминов и т.п. Сведения об определении этих параметров в рамках задачи генерации КТЗ приведены в работе [4]. Перспективным способом классификации предложений после определения параметров можно считать машинное обучение (тогда следует применять схему, аналогичную показанной на рис. 2).

Генерация КТЗ. Авторами разработаны различные алгоритмы генерации КТЗ, отличающиеся глубиной анализа естественного языка. На входе алгоритмы получают начинающиеся с заглавной буквы и оканчивающиеся точкой утвердительные предложения.

Основная проблема при генерации – выбор ключа. Применение случайного выбора ключа на уровне токенов неприемлемо по семантическим требованиям, поэтому выбор ограничивается такими условиями, как: вхождение конкретных слов, конструкций, именованных сущностей, синтаксическая роль анализируемой подстроки в предложении и ее положение в семантическом графе.

Полученные пары источников и ключей (s , k) можно анализировать с точки зрения пригодности методами МО, используя в качестве наблюдаемых переменных данные, полученные от средств ОЕЯ. Гипотеза анализа состоит в том, что смысловая ценность возможного ключа может зависеть от лингвистических характеристик составляющих его слов.

Алгоритмы генерации разделены на две группы. Алгоритмы первой группы основаны на простых предположениях при выборе ключа: они осуществляют просмотр предложения с поиском слов или конструкций определенного вида и заменяют найденную последовательность на пропуски (поиск аббревиатур, числовых значений, определений, конструкций «если ..., то ...»). Алгоритмы второй группы выполняют построение вопроса по результатам морфосинтаксического анализа предложения (вопросы к подлежащему – *что?*, *кто?*; вопросы к прилагательным – *какой(-ая, -ое, -ие)?*; вопросы к обстоятельству места – *где?*; вопросы к обстоятельству времени – *когда?*). Описание алгоритмов и примеры приведены в работах [7], [9].

Оценка качества получаемых КТЗ и их фильтрация. Для реализации схемы фильтрации, показанной на рис. 2, построена статистическая модель логистической регрессии, позволяющая предсказывать вероятность возникновения некоторого события путем подгонки данных к логистической кривой. В задачах классификации объект x относится к классу $y = 1$, если предсказанная моделью вероятность $P\{y = 1|x\} > 0,5$, и к классу $y = 0$ в противном случае. На выходе модель дает вероятность $P(\text{КЛАСС}|\text{ПРИЗНАКИ})$. В данном случае это будет вероятность того, что сгенерированное КТЗ получит положительную оценку на основании присущих ему признаков.

Для выбора ключей рассматриваются лексические признаки (вхождения в потенциальный ключ аббревиатур, слов в верхнем регистре, слов из тематического глоссария, стоп-слов), морфологические признаки (части речи и словоформы), синтаксические признаки (глубина нахождения ключа в синтаксическом дереве, положение ключа относительно сказуемого и подлежащего, часть речи до и после ключа), семантические признаки (положение в семантическом графе).

Для экспериментального исследования метода разработан прототип программного комплекса, содержащий генератор КТЗ и классификатор КТЗ по качеству, основанный на МО.

Литература

1. *Апресян, Ю.Д.* Идеи и методы современной структурной лингвистики / Ю.Д. Апресян. – М., 1966.
2. *Бонч-Осмоловская, А.А.* Разрешение анафоры / А.А. Бонч-Осмоловская, С.Ю. Толдова // Фонд знаний «Ломоносов». – URL: www.lomonosov-fund.ru/enc/ru/encyclopedia:0127469.
3. *Воронец, И.В.* Универсальная автоматизированная система тестирования знаний и самообучения, основанная на анализе естественно-языковых текстов учебных пособий / И.В. Воронец, А.Н. Швецов, В.С. Алешин // Пилотируемые полеты в космос: Сб. докл. V Междунар. науч.-практ. конф. 9–10 апреля 2003 г. – Звездный городок Моск. обл., 2003. – С. 65–67.
4. *Куртасов, А.М.* Возможности применения автоматического реферирования для фильтрации текста в автоматизированной системе генерации тестов из учебных пособий / А.М. Куртасов // Информатизация процессов формирования открытых систем на основе СУБД, САПР, АСНИ и систем искусственного интеллекта: Материалы VII Межд. научно-техн. конф. – Вологда, 2013. – С. 119–122.
5. Методология создания агентно-ориентированных учебных комплексов для подготовки специалистов технического профиля: отчет о НИР (промежуточ.) / Вологодский государственный технический университет; рук. А.Н. Швецов; исполн.: В.А. Горбунов [и др.]. – М., 2011. – № ГР 01201056386.
6. *Танченко, С.С.* Анализ методов генерации тестовых заданий / С.С. Танченко, С.В. Титенко, А.А. Гагарин // XIII Международная научная конференция имени Т.А. Таран «Интеллектуальный анализ информации ИАИ-2013», Киев, 15–17 мая 2013 г.: Сб. тр. / гл. ред. С.В. Сирота. – К., 2013. – С. 220–226.
7. *Швецов, А.Н.* Семантический подход к задаче генерации учебных тестов на основе извлечения знаний из ес-

тественно-языковых текстов / А.Н. Швецов, А.М. Куртасов // Вестник Череповецкого государственного университета. – 2012. – № 1(36). – Т. 2. – С. 11–15.

8. Heilman, M. Good Question! Statistical Ranking for Question Generation. / M. Heilman, N.A. Smith // Human Language Technologies: The 2010 Annual Conference of NAACL. – Los Angeles, California, 2010. – P. 609–617.

9. Kurtasov, A.M. A System for Generating Cloze Test Items from Russian-Language Text / A.M. Kurtasov // Proceedings of the Student Research Workshop associated with

The 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013). – Hissar, Bulgaria, 2013.

10. Mitkov, R. A computer-aided environment for generating multiple-choice test items / R. Mitkov, Le An Ha, N. Karamanis // Natural Language Engineering. – 2006. – 12(2). – P. 1–18.

11. Ng, Andrew Supervised Learning // CS229 Lecture Notes, Stanford University. – URL: <http://cs229.stanford.edu/notes/cs229-notes1.pdf>

УДК 004.8

А.В. Никуличев

Научный руководитель: кандидат технических наук, доцент А.Н. Сорокин

АГЕНТНЫЙ ПОДХОД С ПРИМЕНЕНИЕМ НЕЧЕТКИХ МНОЖЕСТВ ДЛЯ ЗАДАЧИ СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТА

Статья посвящена применению агентно-ориентированного подхода к автоматизации синтаксического анализа предложения на естественном языке. Предлагается методика построения системы, в которой с каждой лексемой связан интеллектуальный агент-исполнитель, сам вычисляющий родительскую связь. За основу рационального выбора альтернативы агентом взят аппарат нечетких множеств, который позволяет использовать интуитивные экспертные знания.

Принятие решений, нечеткие множества, синтаксический анализ, мультиагентная система.

The article is devoted to the agent-based approach to automate text parsing. The technique of constructing the system where each lexical unit is associated with an intelligent agent-performer, calculating a parent relationship, is suggested. The basis for the rational selection of an alternative agent is the formalism of fuzzy sets, which allows the use of intuitive knowledge.

Decision-making, fuzzy sets, text parsing, multiagent system.

В связи с возникновением парадигмы агентно-ориентированных систем, использующих агентов как высокоуровневую абстракцию для формализации и структурирования предметной области и как мощное программное средство для разработки и реализации сложных информационных систем, решение многих задач стало более структурированным и «красивым». В рамках данной статьи будет рассмотрен вариант принятия решений интеллектуальным агентом на основе нечетких множеств применительно к задаче синтаксического анализа.

Следует отметить, что попытки создания автоматического синтаксического анализа применялись уже с 1950-х гг. [3]. Среди всех из них можно выделить анализаторы, использующие оптимально покрытие синтаксической структуры фразы набором правил, описывающих элементарные синтагматические отношения между словами текста [2]. Основная проблема данного подхода – это неоднозначность и сложность составления корректной базы синтагм для последовательной проверки. Таким образом, задача до сих пор является актуальной и требует другого подхода для решения; на смену процедурным алгоритмам пришли алгоритмы, связанные с искусственным интеллектом.

Целью работы является проработка методики синтаксического анализа естественно-языковых предложений на основе мультиагентной системы.

Рассмотрим классическую схему распределенного решения задач несколькими агентами [5] применительно к синтаксическому анализу более подробно:

1. Агент-субординатор (руководитель, центральный орган) проводит декомпозицию исходной проблемы. В рамках нашей задачи он занимается разбивкой предложения на естественном языке на отдельные лексемы и находит их параметры (род, число, падеж и т.д.). Важно отметить, что параметры лексемы не всегда могут быть однозначно определены без контекста и установившихся связей.

2. Отдельные задачи распределяются между агентами-исполнителями. В нашем случае отдельной задачей является нахождение агентом родителя для назначенной ему лексемы. Причем поиск должен осуществляться на основе экспертных знаний, т.е. выбор должен быть обоснованным, и агент должен выбрать наиболее подходящую для него альтернативу.

3. Для получения общего результата производится композиция, интеграция частных результатов, соответствующих выделенным задачам. На данном этапе агент-субординатор также производит анализ результата и корректировку своих решений относительно параметров лексем. Процесс начинается заново, пока не останется вариантов для корректировки. После этого агент-субординатор выбирает самый подходящий вариант среди полученных в результате