электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51036. ISSN 2307-0595

Анализ методов автоматической генерации вопросов на естественном языке

 # 12, декабрь 2015

Тарасенко С. В.[1,*], Рязанова Н. Ю.

УДК 004.5:378.14

[1]Россия, МГТУ им. Н.Э. Баумана
*s.v.tarasenko@ya.ru

Введение

Автоматическая генерация вопросов на естественном языке является одной из актуальных задач компьютерной лингвистики [1]. Системы, обладающие подобной функциональностью, как правило, используются в сфере образования для проверки знаний учащихся, а именно при составлении вопросов по теоретическому материалу [2].

Система автоматической генерации вопросов на основе определенного набора документов должна сгенерировать к ним наиболее полный список вопросов и вариантов ответов. При этом основная трудность заключается в составлении сложных по структуре вопросов [3]. Как правило, такие вопросы основываются на информации из различных частей текста, которая связана между собой только по смыслу. Существующие системы не позволяют составлять сложные вопросы, так как данная задача влечет за собой множество сложностей и проблем, которых можно избежать при составлении простых вопросов из нескольких слов и на основе одного предложения.

Для построения системы генерации сложных вопросов необходимо провести детальный анализ существующих методов генерации вопросов и выбрать из них в наибольшей степени удовлетворяющие основным требованиям поставленной задачи.

1. Генерация вопросов на основе деревьев И/ИЛИ

Метод генерации вопросов на основе деревьев И/ИЛИ позволяет автоматически генерировать набор тестовых вопросов заданного типа. Он широко используется при составлении задач для учебников [4]. Его особенностью является использование уже имеющихся данных о структуре вопроса, что позволяет составлять сложные вопросы, состоящие из неограниченного числа слов и предложений. Метод включает в себя следующие этапы:

1. Построение дерева И/ИЛИ, описывающего структуру вопроса. Все вершины дерева помечаются как постоянные или переменные. В постоянных узлах содержится неизменная информация, а для переменных задаются множества возможных значений.

2. Обход дерева с целью перебора всех возможных комбинаций его узлов. К каждой комбинации применяются определенные алгоритмы, проверяющие ее корректность, т.е. использование соответствующих смыслу вопроса числовых значений и качественных характеристик. В случае если комбинация таковой не является, то при построении вопроса она не учитывается.
3. Формирование набора тестовых вопросов из полученных комбинаций вершин.

Таблица 1. Достоинства и недостатки метода генерации вопросов на основе деревьев И/ИЛИ.

Недостатки	Достоинства
1. Каждый тип вопросов требует отдельного построения дерева; 2. Формирование структуры дерева вопроса, а так же заполнение его узлов (постоянных и переменных) выполняется вручную; 3. Составление алгоритмов для проверки корректности вопросов выполняется вручную для каждого типа вопросов [5].	1. Высокая скорость генерации вопросов; 2. Большое число вариантов тестовых вопросов одного типа.

2. Генерация тестовых вопросов по шаблонам

Метод генерации тестовых вопросов по шаблонам позволяет автоматически генерировать вопросы, используя определенный набор шаблонов [6]. Применение такого шаблона к определенной части текста позволяет выполнить ее перестроение в вопрос. Как правило, данный метод используется для построения вопросов к текстам, находящимся в структурированном формате, таким как словари, представленные в виде массива пар:<Термин, Определение>. Данный метод позволяет формировать вопросы довольно специфических типов: вопрос-меню, заполнение пропущенного поля или указание правильного ответа для его заполнения, расстановка перечисленных элементов в пропущенные поля предложения и т.д.

Таблица 2. Достоинства и недостатки метода генерации тестовых вопросов по шаблонам

Недостатки	Достоинства
1. Ограничение количества типов генерируемых вопросов по причине использования ограниченного числа шаблонов; 2. Возможность генерации вопросов только к текстовым данным в определенном формате [7].	1. Высокая скорость генерации вопроса; 2. Составление вопросов различных типов.

3. Генерация вопросов путем перестроения предложения

Суть метода генерации вопросов путем перестроения предложения заключается в перестроении утвердительного предложения в вопросительное [1]. Он включает в себя следующие этапы:

1. Выбор объекта вопроса. Как правило, вопрос задается к подлежащему, сказуемому, определению или обстоятельству.
2. Подбор вопросительного слова.

3. Перестроение предложения в зависимости от объекта вопроса.

Таблица 3. Достоинства и недостатки метода генерации вопросов путем перестроения предложения

Недостатки	Достоинства
1. Ошибки в определении рода и падежа вопросительного слова; 2. Генерация вопросов на основе только одного предложения; 3. Ограниченное число объектов вопроса [8].	1. Простая реализация.

4. Генерация вопросов на основе текстового корпуса

Метод генерации вопросов на основе текстового корпуса позволяет автоматически генерировать набор различных вопросов сразу ко всему текстовому корпусу, т.е. набору текстов, объединенных определенной общей тематикой [9]. Он включает в себя три основных этапа (рисунок 1):

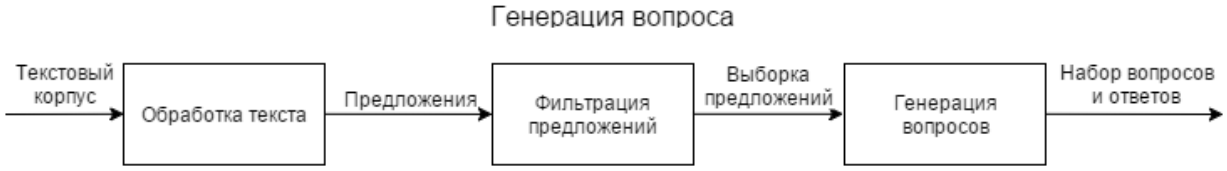

Рис. 1. Основные этапы метода генерации вопроса на основе текстового корпуса

1. Предварительная обработка текста. На данном этапе выполняются подготовительные действия по обработке текста, необходимые для проведения двух последующих этапов:
   1.1. Разбиение текста на предложения. Основная сложность на данном этапе заключается в определении границ предложения. Анализируемый текст может содержать различные аббревиатуры и сокращения, что существенно затрудняет поиск границ предложения. Для решения данной проблемы используются различные регулярные выражения и правила, согласно которым шаблоны применяются к тексту.
   1.2. Построение дерева предложения. Заключается в перестроении предложения из своей линейной формы в древовидную [9]. Полученное дерево размечается стандартным набором тегов [9], которые в дальнейшем используются в процессе формирования вопроса.
   1.3. Дробление сложных предложений. Дерево предложения позволяет с легкостью определить — описывает оно простое предложение или сложное. Суть данного этапа заключается в разбиении сложных предложений на простые, что позволяет упростить дальнейшую обработку текста, поделив его на атомарные структуры.
   1.4. Перерасчет тегов. Данный процесс необходим так, как при перестроении деревьев значения определенных тегов могло поменяться.
2. Фильтрация предложений. Процесс основывается на идее систем автореферирования, которые удаляют из текста «неважные» в определенном контексте предложе-

ния. «Значимость» предложения зависит от количества набранных им очков. На основе их количества удаляются те предложения, число очков которых не вошло в заданный диапазон. При этом необходимо, чтобы каждое предложение было оценено ровно один раз. Очки предложения рассчитываются как среднее среди очков всех входящих в него слов, очки которых обуславливаются их частотой вхождений в текстовый корпус. Далее все предложения сортируются по количеству набранных очков и удаляются предложения с наибольшими и наименьшими баллами.

3. Генерация вопроса. На данном решаются следующие задачи:
   - 3.1. Подготовка предложения. Чтобы облегчить процесс перестроения предложения, на данном этапе опускаются или заменяются некоторые части предложения, представляющие неважные детали. К ним относятся: замена сокращений на соответствующие длинные формы; удаление знаков препинания; написание первого слова в предложении с маленькой буквы, если оно не является именем собственным.
   - 3.2. Определение объекта вопроса.
   - 3.3. Разметка дерева. Помечаются все части дерева, которые в процессе построения вопроса должны оставаться неизменными. Это помогает избежать употребления в вопросе лишней информации.
   - 3.4. Выделение возможных ответов на поставленный вопрос. Заключается в нахождении части предложения, которая может являться ответом на поставленный вопрос. Вариантов ответа может быть несколько. Далее они размечаются аналогично разметке деревьев. Каждому полученному дереву для данного предложения соответствует уникальный ответ. В данном процессе можно задать порядка шести типов вопросов
   - 3.5. Перестановка слов в предложении.
   - 3.6. Генерация ответа. Он известен, так как на основе него мы генерировали вопрос.
   - 3.7. Добавление в начало и конец вопроса вопросительных слов.
   - 3.8. Удаление ответа из дерева.

Таблица 4. Достоинства и недостатки метода генерации вопросов на основе текстового корпуса

Недостатки	Достоинства
1. Сложности в реализации; 2. Невысокая производительность; 3. Использует стороннее ПО [10].	1. Генерирует вопросы к текстовому корпусу; 2. Ошибки при построении вопросов практически отсутствуют.

5. Сравнение методов

Для построения системы генерации сложных вопросов, было проведено детальное сравнение рассмотренных методов по наиболее важным для нее параметрам. Результаты сравнения представлены в таблице 5.

Таблица 5. Пример оформления таблицы

Признак	Генерация вопросов на основе деревьев И/ИЛИ	Генерация вопросов по шаблонам	Генерация вопросов путем перестроения предложения	Генерация вопросов на основе текстового корпуса
максимально возможная длинна вопроса (в словах)	не ограничена	зависит от используемого шаблона	зависит от длинны предложения	не ограничена
количество возможных типов генерируемых вопросов	не ограничено	до 5	1	1
количество членов предложения, к которым может задаваться вопрос	все	подлежащее и сказуемое	подлежащее, сказуемое, определение	все самостоятельные части речи
типичные ошибки при генерации вопросов и их количество	согласование слов при некорректном построении дерева или неверном указании значений в узлах	нет	большое количество неверно согласованных вопросительных слов (по роду, падежу, числу)	нет
зависимость от входных данных	задание структуры дерева, данных в его узлах и правил проверки корректности комбинаций вершин	исходные тексты должны находится в определенном структурированном виде	исходные тексты должны быть разбиты на предложения	нет

Согласно таблице, для составления сложных вопросов разумно использовать метод генерации на основе текстового корпуса. Данный метод позволяет избежать зависимостей от формы представления, объема и размера исходного текста и позволяет генерировать вопросы практически к любому значимому члену предложения. Однако для добавления возможности генерации вопросов различных типов данный метод нужно скомбинировать с методом генерации вопросов по шаблонам, что позволит в большей степени разнообразить конечный набор задаваемых к текстам вопросов.

Заключение

В статье были рассмотрены основные методы генерации вопросов на естественном языке: на основе деревьев И/ИЛИ, по шаблонам, путем перестроения предложения, на основе текстового корпуса. Для каждого метода были подробно описаны его принципы работы, а так же существенные преимущества и недостатки. Для решения задачи построения сложных вопросов было проведено сравнение данных методов по наиболее важным пара-

метрам. В результате было решено использовать метод генерации вопросов на основе текстового корпуса в комбинации с методом генерации вопросов по шаблонам, так как данная комбинация позволяет получать разнообразный набор сложных вопросов различных типов.

Список литературы

[1]. Куртасов А.М., Швецов А.Н. Программа генерации учебных тестов на основе семантического подхода // Труды Международной научно-методической конференции «Информатизация инженерного образования» - ИНФОРИНО-2012. (10—11 апреля 2012 г. Москва). М.: Издательский дом МЭИ. С. 71-74.

[2]. Братчиков И.Л. Генерация тестовых заданий в экспертно-обучающих системах // Вестник Российского университета дружбы народов. Серия: Информатизация образования. 2012. № 2. С. 47-60.

[3]. Максимов В.И., Голубева А.В. Русский язык и культура речи: учебник / под ред. В.И. Максимова, А.В. Голубевой. 2-е изд. СПб.: Златоуст. 2014. 384 с.

[4]. Кручинин В.В. Использование деревьев И/ИЛИ для генерации вопросов и задач // Вестник Томского государственного университета. Томск. 2004. № 284. С. 182-186.

[5]. Кручинин В.В. Использование графов для построения генераторов вопросов в компьютерных учебных программах // Вестник НГТУ. 2004. № 3(18). С. 187-194.

[6]. Кручинин В.В., Морозова Ю.В. Модели генераторов вопросов для компьютерного контроля знаний // Научно-методический журнал. Открытое и дистанционное образование. Томск. 2004. № 2. С. 36-42.

[7]. Сулейманов Д.Ш., Аюпов М.М., Невзорова О.А., Прокопьев Н.А. Семантические технологии генерации учебных вопросов. // Четырнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2014. (24-27 сентября 2014 г. Казань). Труды конференции. Казань: Школа. 2014. Т.3. С. 84-93.

[8]. Yao X., Bouma G., Zhang Y. Semantics-based question generation and implementation // Dialogue & Discourse. 2012. Vol. 3. Is. 2. P. 11-42.

[9]. Aquino J.F., Chua D.D., Kabiling R.K., Pingco J.N. Text2Test: Question Generator Utilizing Information Abstraction Techniques and Question Generation Methods for Narrative and Declarative Text // Proceedings of the 8th National Natural Language Processing Research Symposium. Manila. 2011. P. 29-34.

[10]. Heilman M., Smith N.A. Good question! Statistical ranking for question generation // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. (Proceedings, June 2-4, 2010, Los Angeles, California, USA). Los Angeles: HLT-NAACL. 2010. P. 609-617.