УДК 004.9

АНАЛИЗ НЕКОТОРЫХ АЛГОРИТМИЧЕСКИХ ПРИНЦИПОВ ТРАНСФОРМАЦИИ ТАБЛИЧНОГО РЕФЕРАТА В ТЕКСТ

А.Р. Ятимова, Д.В. Личаргин
научный руководитель канд. техн. наук Личаргин Д.В.
Сибирский федеральный университет

Данная работа посвящена анализу некоторых алгоритмических принципов трансформации табличного реферата в текст. Описываются основные синтаксические положения, позволяющие генерировать текст на основе имеющегося табличного реферата. В настоящее время проблема корректного отображения табличного реферата в текст на основе лингвистических трансформаций является актуальной. Для решения этой проблемы разработаны некоторые принципы, позволяющие выработать чёткую последовательность действий для получения корректного текста в результате развёртки фактологической информации табличного реферата. Различными исследователями ведутся работы в области лингвистики, семантики и искусственного интеллекта. На фоне имеющихся исследований необходимо провести анализ существующих типов, что должно позволить в дальнейшем улучшить автоматические системы трансформации табличных рефератов в текст и реализацию трансформаций вида: «фрагмент базы данных – текст».

Табличный реферат и способы трансформации реферата в текст. Если необходимо придать документальной информации сопоставимую форму, повысить наглядность (если в реферируемом документе соответствующие данные, особенно статистические или параметрические, разбросаны или просто перечисляются в тексте), то, достаточно часто, составляется реферат-таблица. Метод табличного реферата ориентирован на выделение основной информации из источника для облегчения восприятия текста. Основная информация представляется в виде структурированной таблицы с упрощенным содержанием исходного текста. В данном способе описания текста используются логически объединённые строки и столбцы, причём каждый объект текста может быть представлен как строка или столбец матрицы.

Метод табличного реферата выполняет функции структурирования, минимизации и оптимизации текста, а так же концентрации внимания на основных положениях и фактах исходного текста. Табличный реферат может быть использован на широком наборе текстов и для различных целей. Вид конечной таблицы зависит от преследуемой цели. Данный метод применяется не только к научным и техническим статьям, но и к инструкциям и научно-популярным текстам.

Существует множество синтаксических теорий (принципов), благодаря которым можно формализовать получение текста из реферата. Прежде чем описывать существующие принципы, необходимо ввести термин «Трансформация». Трансформация – понятие языкознания, обозначающее то или иное правило, по которому из так называемых ядерных предложений языка (таковыми считаются простые утвердительные предложения с глаголом в изъявительном наклонении активного залога настоящего времени без осложняющих элементов) получаются производные. Возможно применение трансформаций к фразам и предложениям не приведенного вида.

Порождающая грамматика (генеративная грамматика). В конце 1950-х годов Ноамом Хомским был введён термин «Порождающая грамматика» (до этого использовался термин «трансформационная грамматика»). Порождающая грамматика является формализмом генеративной лингвистики, связанным с изучением синтаксиса.

При помощи системы правил, формирующейся в рамках данного подхода, можно определить, какая комбинация слов оформляет грамматически правильное предложение. То есть, в нашем случае, мы имеем набор слов из табличного реферата и можем составить набор правил, который поможет восстановить исходный текст. Но на данном этапе развития грамматики Хомского в несколько большей мере используются для анализа искусственных языков.

Категориальная грамматика. В отличие от порождающих грамматик категориальные грамматики имеют ряд преимуществ в отношении естественных языков. Например, категориальные грамматики позволяют производить семантический разбор параллельно с синтаксическим. Также немаловажным является свойство лексикализации. Оно заключается в том, что вся синтаксическая информация хранится в категориальном словаре и при анализе нужно рассматривать не всю грамматику, а только часть словаря, относящуюся к словам, встречающимся в данном фрагменте текста. В рамках категориальной грамматики каждой синтаксической единице приписывается категориальное значение, или тип. Существует два простых типа: имя (N) и предложение (S), из которых в результате рекурсии могут быть получены сложные типы. Обозначение единицы сложного типа содержит обозначение некоторого более простого типа, а также указание на то, единицей какого типа следует дополнить данную единицу для получения единицы типа S.

Исходя из вышесказанного, можно сделать вывод, что порождающие и категориальные грамматики достаточно хорошо подходят для трансформации табличного реферата в текст.

Теория «Смысл⇔Текст». Данная теория создана И.А. Мельчуком. Она представляет язык в виде многоуровневой модели двух видов преобразований: преобразование смысла в текст и текста в смысл, потому имеет место двойная стрелка в названии теории. Значительная роль отводится толково-комбинаторному словарю – лексическому компоненту модели. Также отличительная особенность это теории – использование синтаксиса зависимостей.

Теория «Смысл ⇔ Текст» – это описание естественного языка, который понимается как устройство («система правил»), обеспечивающее человеку переход от смысла к тексту (построение текста) или от текста к смыслу (интерпретация текста). Но при исследовании языка приоритет отдаётся переходу от смысла к тексту. Согласно данной теории, описание процесса интерпретации текста может быть получено на основе процесса построения текста.

Согласно данной теории построение текста на основе заданного смысла происходит с помощью серии переходов от одного уровня представления к другому, т.е. постулируется многоуровневая модель языка, так как переход осуществляется опосредовано. В настоящее время у данной теории не очень много последователей, и интерес молодого поколения лингвистов к ней не очень значителен.

Метод Филиппа Паркера. Филипп Паркер является профессором и заведующим кафедры экономики в международной бизнес-школе INSEAD. Он создал программу, которая на основе обширной электронной базы может сформировать книгу по заданной теме за достаточно короткое время. Однако данная программа может использоваться только для формирования научных и энциклопедических книг (художественная литератора не поддаётся иерархиям в рамках этого подхода). Таким образом, этот метод может успешно применяться в целях построения табличного реферата, на примере простых шаблонов и их «распаковки».

Существует иерархическая база. Данные в ней хранятся в виде областей, разделов, подразделов, абзацев, предложений и, наконец, отдельных слов. То есть, мы

можем задать определённую тематику и, благодаря такому устройству базы, получить текст из набора ключевых слов.

Рассмотрим пример такого подхода. Мы имеем табличный реферат (см. Таблицу 1). Для восстановления предложения мы можем воспользоваться некоторыми фразами-клише (Summary of the Paper) или схемой изложения краткого пересказа текста (Gist). Согласно методу Филиппа Паркера мы по принципу иерархии извлекаем словосочетания (слова) из таблицы, попутно пользуясь стандартными фразами. Например, Amazon Web Services has been founded in 2002. «Amazon Web Services» является подразделом раздела «Cloud Computing». «…has been founded…» - связка. И «…in 2002.» – снова подраздел раздела, например, «Date». Мы получили простейшее предложение из таблицы. Для добавления более детальной информации необходимо углубленное изучение алгоритма.

Чтобы автоматизировать процесс составления текста из реферата желательно ввести некоторую индексацию элементов таблицы (тогда её можно рассматривать как некоторую матрицу) и производить различные операции над индексированными элементами. В результате можно получить вполне развернутое изложение исходного текста из табличного реферата.

Ниже приводится общая схема изложения краткого пересказа текста на английском языке:

0. Let me <tell you about / retell you> <the fragment / part / paragraph / passage> of the <text / article / book / work>.

1. The main idea of the <text / fragment / ...> is that ..... или The <text / passage> is dedicated to the <idea / point / fact> that .....

2. In particular, … and … <as well as / along with> … are <discussed / analyzed / concerned / considered / touched upon> here.

3. <They are / it is> <viewed / considered / discussed> from the <point of / in the context of / within / in the paradigm of / in the frame(work) of> …

4. That is all I wanted to <tell you / say to you> <about / of / on> the <fragment / passage / text>.

Далее рассматривается схема реферата, с пропусками для вставки слов из табличного реферата на английском языке:

The article / paper / work / publication is dedicated to the topic of …

The main idea of the article / paper / work / publication / text / book / monograph is ...

The … of … such as / like … used in / applied for … within …, for example / for instance / particularly / in particular / especially ... is considered / is viewed / is analyzed / is touched upon …

Specifically / In detail / Partially it is viewed / considered form the point of ..., ... and ... as well as ..., ..., and ...

In the beginning of the text / in the first passage / fragment it is proved / said / stated that the ... includes / refers to / regards / concerns / ... the ...

Further / next / after that / afterwards … is viewed / is offered / is given / is presented / is shown / is exhibited / is demonstrated.

As …, the … of … is viewed / …

In conclusion / finally / to sum it up / as a generalization / all in all the author describes / views / determines / defines / specifies / analyzes / touches upon | the ... of ... within ... during ... including ... used in ... based on ... aimed at ....

Ниже рассматривается пример текста, который может быть построен из следующего табличного реферата (см. Таблицу 1).

The publications are dedicated to the topic of Cloud computing. Cloud services are computing services offered by a third party, available for use when needed, that can be scaled dynamically in response to changing needs. The main idea of this publication is about Cloud computing techniques, reasons and benefits. The first commercial cloud service was developed by Amazon company. Cloud services help to solve the problem with lack of hardware resources without buying own servers and creating own data center. Usage of suggested services gives enormous economical benefit, scalable hardware resources. Five main principles that define cloud computing are … Cloud computing is a very useful technology, which is important in evolution of IT. It gives many new opportunities for IT business.

Таблица 1

Табличный реферат на тему облачные вычисления

Objects / **Aspects**	**Cloud computing …**		
	Infrastructure as a Service (IaaS)	**Platform as a Service (PaaS)**	**Software as a Service (SaaS)**
Vendors	Amazon Web Services Sun Grid Engine	Microsoft Azure Google App Engine	GRIDS Lab Force.com
Has been founded in	2002 2004	2006 2008	2008 2009
Security	SAS 70 Type II Certification		
Applications	Has different / various / numerous	-	+
Data	-	-	+
Programming Framework	-	.net Python, Java	C, C# Apex
OS	Maybe shared	+	+
Virtualization	+	+	+
Servers	+	+	+
Storage	+	+	+

Важно отметить, что один и тот же табличный реферат может быть преобразован во множество различных текстов на естественном языке. В целях обеспечения вариативности генерации подобных текстов необходимо привлечение следующих понятий и концепций: семантические классификации слов и понятий естественного языка, базы правил трансформаций слов и конструкций языка, модули учета частотных и стилистических характеристик генерируемых фраз языка и др.

В заключение необходимо отметить, что проведённый анализ некоторых принципов трансформации, может быть, применим к табличному реферату. Категориальные грамматики и метод Филиппа Паркера наиболее применимы к данному типу рефератов.