УДК 004.65:81`322

ГЕНЕРАЦИЯ ПОДМНОЖЕСТВ ЕСТЕСТВЕННОГО ЯЗЫКА НА ОСНОВЕ ГИБРИДИЗАЦИИ ПОРОЖДАЮЩИХ ГРАММАТИК И МНОГОМЕРНЫХ БАЗ ДАННЫХ

Личаргин Д.В., Усова А.А., Ладе А.В.

ФГАОУ ВПО «Сибирский федеральный университет», Красноярск, e-mail: lichdv@hotmail.ru

Цель работы состоит в необходимости сформулировать некоторые принципы генерации фраз естественного языка общей тематики в целях повышения эффективности процесса составления учебных материалов лингвистическим программным обеспечением. Рассматривается проблема усовершенствования метода генерации учебных заданий на основе многомерного анализа семантических данных за счет привлечения алгоритма автоматической генерации используемых шаблонов на основе порождающих грамматик. Проблема разработки все больших объемов учебных материалов является актуальной в связи с внедрением индивидуальных подходов и траекторий обучения. Для решения этой проблемы предлагается использовать гибридный метод работы с многомерным представлением семантически векторизованных данных на основе порождающих грамматик. Разрабатывается программа генерации осмысленных фраз естественного языка общей тематики, а также автоматической генерации учебных материалов на основе предложенных подходов. Делается вывод о необходимости дальнейших исследований в области генерации естественного языка на основе генерируемых семантических шаблонов над многомерными массивами данных.

Ключевые слова: компьютерная лингвистика, генерация учебных заданий, генерация осмысленной речи

GENERATION OF NATURAL LANGUAGE SUBSETS BASED ON HYBRIDIZATION OF GENERATIVE GRAMMARS AND MULTIDIMENSIONAL DATABASES

Lichargin D.V., Usova A.A., Lade A.V.

Siberian Federal University, Krasnoyarsk, e-mail: lichdv@hotmail.ru

The purpose of the work is to formulate some principles of generating a common theme phrases in the natural language to increase the efficiency of the process of educational materials composition by linguistic software. The problem of improving the method of learning tasks generation based on multidimensional analysis of semantic data by using the algorithm of automatic generation of the applied patterns based on generative grammars is considered. The problem of the development of the optionally large amounts of training materials is topical regarding the implementation of individual approaches and learning trajectories. For solving this problem, a hybrid method of working with multidimensional representation of semantically vectorized data and generative grammars is proposed. The software for generating meaningful phrases of a common theme of the natural language as well as automatic generating various educational materials based on the proposed approaches is now developed. The conclusion about the necessity of further research in natural language generation field based on previously generated semantic patterns over multidimensional data arrays is made.

Keywords: computational linguistics, generation of educational tasks, generation of meaningful speech

Проблема генерации осмысленной речи исследуется достаточно давно. Проблема генерации осмысленной речи исследовалась различными авторами, такими как К. Шеннон, Т. Виноград, Э. Кодд, А. Хомский, М.В. Никитин, А.С Нариньяни. Центральными задачами в данной области являются перевод, машинный перевод, построение экспертных систем.

Предлагается модель генерации учебных заданий на основе многомерного анализа семантических данных за счет привлечения алгоритма автоматической генерации используемых шаблонов на основе порождающих грамматик. Предложенная модель позволяет обеспечивать разработку тестовых заданий по различным разделам языка, автоматически генерировать ответы на эти задания, гибко варьировать требуемый формат данных для различных обучающих систем.

Однако вопрос создания лингвистического программного обеспечения на основе логики формальных описаний дискретной математики [1–4] требует дополнительных исследований, в частности, с привлечением методов генерации осмысленных высказываний на основе подстановочных таблиц, метода семантической классификации, метода векторизации многомерных данных.

Актуальность разработки системы состоит в необходимости повышения эффективности работы преподавателей, сокращении временных затрат, увеличении объемов генерируемых тестовых материалов, которые позволят повысить качество обучения за счет использования множества вариантов индивидуальных заданий. Для разработки обучающих материалов от преподавателя требуется привлечение значительных временных ресурсов, что делает процессы методического направления работы пре-

подавателя недостаточно эффективными и результативными. Поэтому необходима разработка ресурса, позволяющего преподавателям иностранных языков генерировать учебные задания в целях проверки знаний учащихся. Таким образом, задача разработки такого ресурса является актуальной. Новизна работы определяется существенным улучшением порождающей мощности формальных грамматик за счет их гибридизации с многомерными массивами данных.

Цель данной работы заключается в разработке программы повышения вариативности генерируемых фраз естественного языка, на основе гибридизации порождающих грамматик и многомерных баз данных.

Задачи работы, рассматриваемой в данной статье, состоят:

1) в разработке порождающей грамматики по генерации грамматико-семантических шаблонов порождения осмысленной речи;

2) разработке многомерных баз данных на основе авторской семантической базы данных;

3) разработке модели, позволяющей обеспечить гибридизацию многомерных баз данных и порождающих грамматик.

Основная идея решения проблемы недостаточной порождающей мощности формальных грамматик состоит в использовании принципа гибридизации многомерных баз данных и порождаемых шаблонов. Порождаемые шаблоны помогут сделать задания менее стандартными с точки зрения пользователя, а также разнообразить их содержание. Использование генераторов учебных заданий может со временем стать все более привлекательным, что сделает использование или редактирование чужих материалов с возможным нарушением авторских прав неприемлемым и нецелесообразным не только с этической и юридической, но и с технической точки зрения.

Рассмотрим многомерное пространство единиц естественного языка, в частности слов и предложений как математического объекта, составляющего критерий осмысленности генерируемых фраз естественного языка. Несоответствие грамматическим нормам безотносительно осмысленности семантического значения дают фразы вида «See I», «Love time space» и тому подобные строки случайных слов. Соответствие грамматическим нормам при отсутствии осмысленности семантического значения дают фраза «I eat a hat», «Friendship makes breakfast» и тому подобные бессмысленные фразы, имеющие грамматически корректную синтаксическую структуру. Для задания соответствий между словами, определяющих грамматическую и семантическую осмысленность порождаемой письменной речи, необходимо использовать поэтапно верифицируемый критерий осмысленности генерируемых фраз. Таким критерием могут служить подстановочные таблицы, позволяющие генерировать более 98 % осмысленных фраз от всех возможных сочетаний слов из подставляемых в синтагматические отношения групп слов, соответствующих колонкам подстановочных таблиц.

Ниже приводится пример учета комбинаторики слов естественного языка, представленного в форме подстановочной таблицы, позволяющей генерировать осмысленные фразы на английском языке.

Такого рода шаблоны в форме реляционных таблиц лингвистических данных обеспечивают генерацию фраз вида: «the user stops optimizing the program», «the cracker continues improving the file» и прочее.

Такие шаблоны образуют срез лексико-грамматического пространства слов естественного языка. Логика генерации осмысленных фраз естественного языка может быть спроецирована и быть эквивалентной логике порождающих грамматик. Само лексико-грамматическое пространство слов естественного языка описывается как минимум тремя основными координатами.

Координаты многомерного лексико-грамматического пространства соответствуют элементам вектора признаков классификации слов языка, этими координатами являются следующие:

1. Порядок слов естественного языка вида: Делатель – Действие – Предмет –Реципиент – Место – Время – Причина – …

2. Темы генерируемых фраз естественного языка: Еда, Одежда, Компьютеры, Здания, Печатные издания, …

3. Варианты подстановок слов по одной из тем: Повар, кок – готовит, жарит, тушит – овощи, репу – на кухне, в кафе – в обед, в пять часов – по запросу клиента, по распоряжению начальства…

Данное лексико-грамматическое пространство предполагается гибридизировать с системой порождающих грамматик для более эффективной обработки строк естественного языка.

Порождающие грамматики помогут осуществить порождение шаблонов генерации, которые будут ссылаться на массивы ячеек многомерных баз данных. На основе шаблонов, порождаемых формальной грамматикой, программа генерирует фразы, имеющие вид: «This responsible producer invents various keyboards. This responsible user sets up a complex computer. That clever woman makes a good project. This clever researcher purchases a large desktop».

Срез многомерного пространства слов языка в виде подстановочной таблицы

Зэ ... ***этот ...***	...-(e)s ...-(и)с/з **...**	...-ing ...-иН **...**	the ... Зэ ... ***этот ...***
cracker ***взломщик программного обеспечения***	finish ***заканчивать***	optimize ***оптимизировать***	software ***программное обеспечение***
user ***пользователь***	stop ***завершить***	improve ***улучшать***	file ***файл***
private user ***частный пользователь***	continue ***продолжать***	maintain ***поддерживать в хорошем состоянии***	project ***проект***
client ***клиент***	control ***контролировать***	make an error in ***допускать ошибку в***	program ***программа***

В программе «Генератор учебных заданий на основе многомерных данных» можно создавать задания для занятий по английскому языку. В связи с этим в программе возможно использование промежуточного языка типа Интерлингва. То есть сначала фразы генерируются на промежуточном языке [5], а потом строка таких нетерминальных символов заменяется на терминальные символы естественного языка.

Внедрение промежуточного языка и гибридизация многомерных баз данных и порождающих грамматик поможет повысить порождающую мощность формальных грамматик, снизить процент недостаточно естественно звучащих фраз генерируемого подмножества естественного языка.

Предполагается, что рассматриваемое представление данных будет давать возможность программной системе сгенерировать условно произвольное число вариантов заданий по иностранному языку.

Дерево состояний абстрактной строки подмножества естественного языка, генерируемое на основе порождающих грамматик с правилами, ссылающимися на разделы многомерных баз данных, визуализируется в рамках интерфейса разрабатываемой программной системы. В программной системе «Генератор учебных заданий на основе многомерных данных» реализуется модель, позволяющая получать на выходе для конечного пользователя учебные материалы с привлечением лингвистических многомерных и сопутствующих баз данных на входе системы. Эта модель данных дает возможность реализовать функционал разрабатываемой программной системы на основе алгоритмов, в основе которых лежат представления о системе естественного языка.

Эта модель данных является отражением семантической структуры множества слов языка и внутренней структуры формального описания смысла слов естественного языка, что обеспечивает возможность выделения множества осмысленных фраз из множества бессмысленных с использованием критериев осмысленности. В частности, в работах таких ученых семасиологов, как М.В. Никитин, описываются понятия импликации одной группы слов на основе другой. Например, слово «здание» может имплицировать такие слова, как «кирпичное», «бетонное», «многоэтажное», «жилое», «комфортабельное». При наличии не двух, а нескольких аналогичных групп слов накладываются ограничения на множество осмысленных, частотных, естественно звучащих фраз, порождаемых на основе анализа такого рода соответствий.

Программная система учитывает частотные принципы семантической организации подмножеств естественного языка, упорядоченных на основе вектора семантической классификации. В частности, разработанная программная система содержит в себе различные клоны одних и тех же групп слов в зависимости от их частотности, в результате чего порождаемые фразы имеют вид: «Мой друг съел соленые огурчики» (частотная фраза) в отличие от фраз, генерируемых без учета частотных принципов: «Мой шурин жует постную брокколи» (низкочастотные фразы со стилистическим эффектом смысловой перегруженности, комичности или ошибочности). Аналогично фраза «Разжигатель войны чавкает чрезвычайно размороженной пиццей» является неестественной (не частотная фраза). Вероятность употребления фразы с большей или меньшей частотностью можно в первом приближении описать следующей формулой:

$$P(w_{1\text{-}n}) = P(w_1) * P(w_2) * \ldots * P(w_n),$$

где

$$P(w_i) = 1 / \log_2 (1 / F(w_i)) * q,$$

где P – вероятность употребления строки символов, а F – частотность употребления

слова (единица, деленная на среднее количество слов, на которое приходится одно его употребление) w_i, q – коэффициент важности лексического разнообразия для каждого конкретного стиля речи (официального, поэтического, академического и пр.).

В результате решается проблема не только выделения множества осмысленных фраз из множества любых фраз, включая бессмысленные, но и выделение из множества осмысленных фраз наиболее частотных.

Многомерный анализ текста в программе «Генератор учебных заданий на основе многомерных данных» дает возможность учитывать прецеденты употребления сочетаний слов в текстах, по мотивам которых создаются учебные задания. Такая генерация на основе прецедентов не является процессом десинонимизации и обхода авторских прав, а связана лишь с учетом общего контекста текста прецедентов в качестве учета общей тематики и некоторого общего лексического наполнения для урока иностранных языков.

Предложенная модель (см. рисунок) описания лингвистических данных включает алгоритмический этап для обработки ссылок на многомерные базы данных. В результате решается проблема, состоящая в том, что для генерации фраз осмысленного языка на основе неприемлемо большого набора правил, рассматривающих все более специфические случаи словоупотребления. Соответственно необходимо большое количество человеко-часов работы, и часто подобное решение задач упирается в комбинаторно большое увеличение требуемого количества привлекаемых данных (большие данные), гиперболического роста количества правил, необходимых для генерации определенного подмножества языка. Эта проблема была решена посредством введения в синтаксис правил порождающих грамматик ссылок на классы строк, реализующихся на уровне математической абстракции в виде многомерных кубов семантической базы данных. Сам синтаксис ссылок на раздел многомерных баз данных может варьироваться на основе порождающих грамматик, например можно добавить нужный префикс к ссылке на группу слов, означающий степень частотности этой фразы. Аналогичным образом снижается соотношение количества генерируемых фраз к количеству используемых фраз порождающей грамматики.

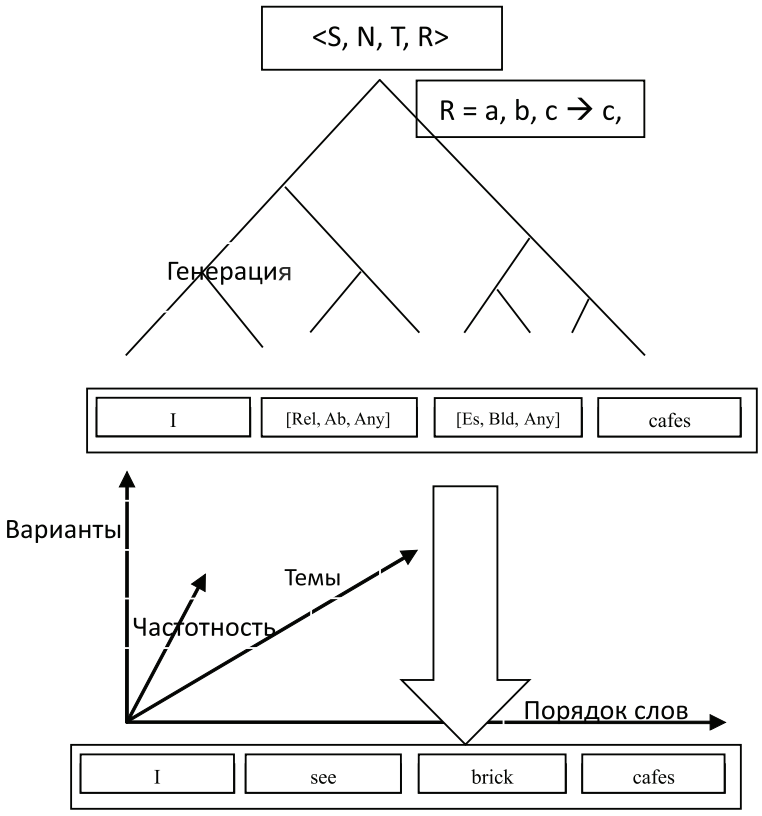

Схема генерации предложений рассматриваемой гибридной модели

В статье рассматривается проблема гибридизации многомерных баз данных и порождающих грамматик в системе генерации осмысленных заданий и фраз естественного языка. Рассматривается реализация этого подхода при разработке и усовершенствовании программы «Генератор учебных заданий на основе многомерных данных», которая будет способствовать автоматизации работы преподавателей по разработке электронных образовательных ресурсов, а также позволит ускорить процесс разработки тестовых заданий на основе элементарных лексико-семантических языковых элементов.

Список литературы

1. Личаргин Д.В. О функции плановых языков на современном этапе и их применении в качестве языков нетерминальных символов порождающих грамматик / Д.В. Личаргин, А.В. Ладе, Д.Д. Мищенко, А.Т. Гордеева // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2014. – № 1 (53). – С. 44–48.

2. Личаргин Д.В., Таранчук Е.А. Иерархическая структура учебного электроного курса и его вариабельность для обучения иностранному язык. // Журнал «Дистанционное и виртуальное обучение». – 2011. – № 4. – С. 56–75.

3. Сафонов К.В., Личаргин Д.В. Некоторые принципы автоматической генерации учебных материалов на основе баз знаний и лингвистической классификации // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2012. – № 2 (42). – С. 72–77.

4. Сафонов К.В., Личаргин Д.В. Разработка векторизованной семантической классификации над словами и понятиями естественного языка // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2009. – № 4 (25). – С. 33–37.

5. Safonov K.V., Lichargin D.V. Elaboration of a vector based semantic classification over the words and notions of the natural language // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2009. – № 5 (26). – С. 52–56.