УДК 004.9

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ЯЗЫКА КАК СПОСОБ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ГЕНЕРАЦИИ ЯЗЫКА ПО ШАБЛОНАМ ИЗ МНОГОМЕРНЫХ БАЗ ДАННЫХ

Маглинец А.Ю., Личаргин Д.В.
научный руководитель канд. техн. наук Личаргин Д.В.
Сибирский Федеральный Университет

Аннотация

В статье рассматривается проблема определения семантического метрического расстояния между фрагментами текста на естественном языке с использованием шаблонов генерации осмысленных фраз языка в интеграции с применением анализа корпусов текстов на основе модели Марковских процессов. Проблема давно рассматривается с точки зрения методов компьютерной лингвистики, искусственного интеллекта, традиционной лингвистики, информатики.

Рассматривается вопрос о методах построения шаблонов генерации осмысленных фраз и различных видов проекций информации из этих шаблонов на предложения из корпусов текстов на естественном языке. Рассматривается также вопрос о проецировании частично бессмысленных фраз и фраз с ошибками на данные шаблоны в рамках проблемы автоматической семантической корректуры текста. Делается вывод о необходимость гибридизации статистических и парадигматических методов генерации осмысленных подмножеств языка в приложении к различным задачам, в частности, устранения семантических ошибок, перевода с элементами реферирования и семантической стандартизации текста.

В работе рассматривается проблема построения алгоритма вычисления вероятности проекции фрагмента текста на семантические шаблоны реляционной базы данных. На сегодняшний день широко распространены и разрабатываются разнообразные системы анализа текстов на естественном языке, используются различные методы и критерии отделения осмысленных фраз языка от бессмысленных, в частности, особо важную роль в современной дисциплине «обработка естественного языка» играет статистический метод определения осмысленности фраз. Будучи наиболее проработанным методом на сегодняшний день, он позволяет формировать достаточно работоспособные модели на основе Марковских процессов. Проблема является актуальной в связи с тем, что анализ и аналитика текстов на естественном языке требует больших человеческих и временных ресурсов и нуждается во всё большей автоматизации. Проблема генерации множества осмысленных фраз языка решается на стыке таких наук, как компьютерная лингвистика, искусственный интеллект, традиционная лингвистика, информатика, психология.

Проблема генерации естественного языка давно и широко исследуется различными авторами, в частности Т. Виноград, Р. Г. Пиотровским, К. Шенноном, А. Тьюрингом и многими другими.

Цель данной работы состоит в описания алгоритма определения соответствия «фрагмент текста – фрагмент базы данных» в первом приближении. Задачи данной работы заключаются в:

1. Разработке и верификации определённого количества шаблонов генерации осмысленного языка по темам вузовского \ школьного английского языка.

2. Проработка алгоритма и мысленный эксперимент с оригинальными англоязычными текстами, относящимися к теме учебника английского языка.

3. Реализация данного алгоритма в виде программного кода, предназначенного для генерации фраз к учебным заданиям по выбранным тематическим текстам.

Основная идея работы состоит в построении гибридной модели численной оценки вероятности вхождения предложения во множество языка с учетом вхождения пар, троек и т.д. слов во множество предложений корпуса текстов. Новизна работы состоит в том, чтобы предложить формулу оценки данной вероятности.

Статистический критерий осмысленности. В работах М. Коллинза, Колумбийский Университет, рассматривается вопрос об оценки условного параметра осмысленности фраз естественного языка на основе статистических оценок словоупотребления в контексте пар, троек, четверок и пятерок слов в корпусах текстов на основе модели Марковских процессов. Осуществляется вероятностная оценка возможности вхождения той или иной фразы во множество осмысленных фраз языка с учётом вероятностных оценок прецедентов вхождения однёрок-двоек-троек и так далее слов в большие по объёму корпусы текстов.

Парадигматический критерий осмысленности. Данный критерий основывается на классификации понятий и слов естественного языка. Для этого используются деревья классификации, узлами которых являются понятия. Каждому понятию может соответствовать одно слово или целый ряд синонимов или же не соответствовать ни одного слова. Набор узлов-понятий назовем понятийным пространством. Каждому уровню дерева соответствует один определенный семантический дифференцирующий признак, с конечным, фиксированным набором значений. Признаки для узлов разных уровней составляют ряд или вектор признаков классификации. Признаки одного ряда имеют тождественные элементы, определяющие связи и логику классификации, выраженную в виде семантической формулы определяемых понятий.

Значения каждого элемента семантического вектора являются понятиями другой классификации – более низкого уровня. Единицы языка разного уровня – предложения, слова и понятия, семы (атомы смысла) и т. д. представлены соответствующими классификациями разного уровня (или как говорят лингвисты разного «яруса»). Каждая классификация генерирует понятия для создания ряда признаков последующих классификаций. Каждую из классификаций задает свой вектор семантических признаков. На основании каждого из этих векторов можно построить семантическую формулу, дерево классификации или многомерное понятийное пространство общие для всех единиц одного уровня.

Признаки классификации. Как указывается в работах таких авторов как К.В. Сафонов, Д.В. Личаргин и др., вектор признаков классификации задается перечислением множества значений элементов семантического вектора или же при помощи порождающей грамматики для каждого уровня классификации, то есть для каждого элемента вектора отдельно. Зададим семантический вектор классификации для слов и понятий естественного языка. В данном семантическом пространстве работает метрика Хэмминга, при этом в некоторых случаях имеет смысл использовать евклидову метрику

В предложенной классификации слов слова разбиваются на классы и подклассы, хорошо сочетающиеся друг с другом комбинаторно и/или ассоциативно. На основе этого принципа разработан электронный словарь, позволяющий генерировать подстановочные таблицы в целях генерации осмысленных фраз и текстов

пользователем или программным обеспечением. Ниже дается пример подобной подстановочной таблицы.

В частности, подстановочная таблица по теме «симпатии к одежде», подстановочная таблица по теме «поход в магазин» и далее - по теме «деньги за товар» образуют последовательность подстановочных таблиц, выборка предложений из которых дает предложения вида: «я люблю полосатые жакеты, я с удовольствием ношу полосатую одежду. Завтра я иду в магазин на улице Иванова. Я еду туда на машине. Я заработал 50 долларов и хочу потратить 300 рублей на новый жакет». Таким образом, два вышеупомянутых уровня классификации не только определяют позицию классов слов в понятийном пространстве, но и могут входить в классификацию фраз, организованную тематически. Последнее должно позволить визуализировать в рамках естественно-языкового интерфейса не только структуру предложения, но и структуру возможных текстов.

Таблица 1

Подстановочная таблица как средство генерации осмысленных фраз.

I ***я***	can ***может***	drive ***водить***	my ***мой***	car ***автомобиль***
We ***мы***	could ***мог бы***	ride ***везти***	your ***твой***	bus ***автобус***
You ***вы***	may ***может (с разрешения)***	take ***сесть на***	his ***его***	means of transport ***транспортное средство***
they ***они***	might ***мог бы (с разрешения)***	get on ***сесть на***	her ***ее***	plane ***самолет***
he ***он***	Shall ***следует***	sit in ***сесть в***	our ***наш***	airplane ***аэроплан***

В рамках рассмотрения этих двух различных критериев осмысленности, необходимо отметить, что каждый из рассмотренных критериев не является достаточным сам по себе для решения задач, связанных с определением семантического метрического расстояния между фрагментами текста на естественном языке. Статистический критерий осмысленности не учитывает семантические аспекты языка, делая определение осмысленности фразы затруднительным. Однако он позволяет легко выявлять узуальные фразы, которые часто встречаются в корпусах текстов.

Парадигматический критерий осмысленности позволяет проводить оценку осмысленных подмножеств языка с точки зрения логической совместимости используемых понятий. Однако полноценная оценка затруднительна ввиду несовершенства алгоритмов и электронных словарей, особенно ярко это несовершенство проявляется на текстах и фразах с высокой окказиональностью.

Также следует заметить, что критерии узуальности и окказиональности могут различаться в зависимости от культурных различий тех или иных наций, их привычек и традиций. Если предложение «я люблю газированную воду» будет одинаково воспринято представителем практически любой нации, то фразу «на завтрак будут жареные огурцы и салат из одуванчиков» представитель европейской культуры, скорее всего, воспримет как имеющую шуточную окраску, в то время как для коренного жителя Китая подобное высказывание будет звучать вполне привычно в силу особенностей китайской национальной кухни.

Совмещение двух критериев определения осмысленности фраз естественного языка, а именно, статистического и парадигматического метода даёт очевидное преимущество.

В то же время есть критерии узуальности \ окказиальности, которые диктуются конкретными задачами и под которые надо приспосабливать разрабатываемые решения и алгоритмы. То есть фразы с очень большой узуальностью не всегда являются удачными, потому что в реальной жизни сложно, точнее даже невозможно встретить носителя языка, который бы говорил на нем на сто процентов узуально.

После рассмотрения методов, приводимых выше, необходимо отметить, что использование гибридизации данных методов открывает новые возможности по анализу текстов и использованию критериев осмысленности фраз естественного языка. При их совмещении возможно создание системы, которая бы генерировала тексты на основе подстановочных таблиц (учитывая при этом семантику многомерной классификации и глубинных индексов семантического значения слов), и после этого оценивала бы их узуальность \ окказиальность употребления фраз языка на основе гибридных оценок, на основе корпусов текстов.

Программная система должна оценивать вероятность встретить то или иное, принципиально и логически возможное предложение с точки зрения его допустимости, привычности и общеупотребительности на основе статистических методов.

Что касается автоматического исправления ошибок, традиционные методы исправления ошибок в целом связаны с анализом грамматических структур на основе порождающих грамматик Хомского. На основе гибридных методов оценки осмысленности текстов возможно предложить пользователю варианты предложений более приведенного вида: например, вместо «я желание понять ты» будет предложен семантический вариант «я хочу понять тебя».

Таким образом, предложим следующую формулу оценки допустимости сгенерированной по шаблонам фразы на основе статистических методов оценки их вхождения в корпус текстов:

$$F\left(\log_2 \frac{q(A)}{h}\right)\cdot F\left(\log_2 \frac{q(B)}{h}\right)\cdot F\left(\log_2 \frac{q(C)}{h}\right)\cdot F'\left(\sum_{i=1(A,B)}^{q(A,B)} \frac{1}{|S_i(A,B)\cdot k - S'(A,B)\cdot k'|+m}\right)\cdot$$
$$F'\left(\sum_{i=1(B,C)}^{q(B,C)} \frac{1}{|S_i(B,C)\cdot k - S'(B,C)\cdot k'|+m}\right)\cdot F'\left(\sum_{i=1(A,C)}^{q(A,C)} \frac{1}{|S_i(A,C)\cdot k - S'(A,C)\cdot k'|+m}\right),$$

где q(a,b) – количество пар слов (a, b), встречаемых на незначительном расстоянии в предложениях корпусов текстов, h – общее количество предложений в корпусе текстов, S_i(a, b) – интервальное расстояние между словами a и b в i-том предложении без учёта однородных членов предложения в корпусе текста, S’(a, b) – расстояние между словами в шаблоне генерации, k – коэффициент для увеличения величины результирующих малых вероятностей, m – коэффициент для исключения деления на ноль.

Необходимо оценка вариантов выбора функций F и F` для определения оптимального распределения вероятностей вхождения предложения во множество языка. Это могут быть такие распределения как:

1) Нормальное;
2) Линейная функция;
3) Степенная функция;
4) Mexican hat;

Выводы. В работе выполнен анализ проблемы гибридизации статистических и парадигматических методов генерации осмысленных подмножеств языка в приложении

к различным задачам. Предложена численная модель оценки вероятности вхождения предложения во множество языка с учетом вхождения пар, троек и т.д. слов во множество предложений корпуса текстов. Рассматриваются возможности применения этих методов для автоматической проверки семантических ошибок в текстах на естественном языке. Подчеркивается важность продолжения исследований по теме определения семантического метрического расстояния между фрагментами текста на естественном языке. Делается вывод о необходимость учета функций распределения вероятности вхождения предложения во множество языка с учетом вхождения пар, троек и т.д. слов во множество предложений корпуса текстов.