# Статистические методы анализа языка и метод генерации языка по шаблонам из многомерных баз данных

Личаргин Дмитрий Викторович
*Сибирский федеральный университет (Красноярск), Россия*

Маглинец Артем Юрьевич
*Сибирский федеральный университет (Красноярск), Россия*

Рыбков Михаил Викторович
*Сибирский федеральный университет (Красноярск), Россия*

## Аннотация

В работе рассматривается проблема построения алгоритма для 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 (Т. Виноград, Р. Г. Пиотровский, К. Шеннон, А. Тьюринг и многие другие).

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 (М. Коллинза, Колумбийский Университет).

Рассматривается векторизованная семантическая классификация слов 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 (Сафонов К.В., Личаргин Д.В. и др.).

Основная идея работы состоит в построении гибридной модели для численной оценки вероятности вхождения определенного предложения во множество шаблонов генерации естественного языка с учетом вхождения пар, троек и т.д. слов во множество предложений корпуса текстов.

В рамках рассмотрения двух различных критериев осмысленности – статистического и парадигматического, необходимо отметить, что каждый из рассмотренных критериев на сегодня 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, однако пол-

ноценная оценка затруднительна ввиду неполноты и несовершенства электронных словарей и баз данных.

Предлагается формула оценки допустимости сгенерированной по шаблонам фразы на основе гибрида статистических и парадигматических методов оценки их вхождения определенных фраз во множество осмысленного языка с учетом как шаблонов генерации фраз естественного языка, так и статистической обработки корпусов текстов.