

УДК 81'33

## ОБЗОР ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ, ИСПОЛЬЗУЕМЫХ В СОВРЕМЕННОЙ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКЕ

**А.О. Варыгин**

*Кубанский государственный университет  
Краснодар, Россия, [andrejvarygin67@gmail.com](mailto:andrejvarygin67@gmail.com)*

В компьютерной лингвистике используют вероятностные модели, например, для автоматической обработки естественного языка. Выбор модели определяется характером задачи (локальный контекст, последовательная разметка, классификация или анализ корпусов), а их комбинация даёт наилучший результат.

*Ключевые слова:* вероятностные модели; компьютерная лингвистика; n-граммы; скрытые марковские модели; теорема Байеса; LDA.

Современная компьютерная лингвистика пытается решить задачи автоматизации больших объёмов текстовой информации. Детерминированные подходы (правила, грамматика, конечные автоматы, регулярные выражения, логические выводы и т.д.) иногда не могут учесть сложность живого общения и для этого используют вероятностные модели, которые оценивают неопределённость и могут предсказывать наиболее вероятные варианты.

Рассмотрим четыре вероятностные модели, которые наиболее распространённые, благодаря сочетанию простоты и эффективности, а также в силу своей универсальности и вычислительной надёжности. Сравнительная характеристика представлена на рисунке 1.

| Сравнительная характеристика вероятностных моделей                |                                                          |                                                         |                                                        |
|-------------------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------|--------------------------------------------------------|
| в компьютерной лингвистике: задачи, сильные стороны, примеры      |                                                          |                                                         |                                                        |
| Модель                                                            | Основная задача                                          | Сильные стороны                                         | Пример использования                                   |
| <b>N-граммные модели</b><br>оценка вероятности слова              | Предсказание следующего слова,<br>языковое моделирование | Простота, высокая скорость<br>Лёгкость обучения         | Автодополнение текста,<br>распознавание речи (ASR)     |
| <b>Скрытые марковские (HMM)</b><br>скрытые состояния → наблюдения | POS-тегирование,<br>анализ последовательностей           | Учёт контекста (переходы)<br>Алгоритм Витерби (оптимум) | Разметка частей речи, NER,<br>токенизация (яп., кит.)  |
| <b>Байесовские модели</b><br>теорема Байеса                       | Классификация текстов,<br>фильтрация спама, тональность  | Работает с малыми данными<br>Высокая интерпретируемость | Спам-фильтры,<br>сентимент-анализ отзывов              |
| <b>Тематическое LDA</b><br>латентные темы                         | Кластеризация документов,<br>выявление скрытых тем       | Работа с большими корпусами<br>Снижение размерности     | Рубрикация научных статей,<br>рекомендательные системы |

**Рисунок 1. – Сравнительная характеристика**

1) N-граммные модели оценивают вероятность появления слова на основе  $N - 1$  предыдущих слов. Простой случай – биграмма, где вероятность слова зависит только от одного предшествующего слова:

$$P(w_i|w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}.$$

Например, в интернет-магазине в поисковой строке можно предсказать следующее слово. По статистике, после слова *дешёвый* часто вводят *телефон*. Тогда:

$$P(\text{телефон}|\text{дешёвый}) = \frac{\text{count}(\text{дешёвый телефон})}{\text{count}(\text{дешёвый})}.$$

Пользователю предлагаются релевантные подсказки уже на втором слове. Плюс – мгновенный отклик. Минус – если пользователь написал *недорогой и качественный...*, модель будет ориентироваться только на последнее слово (*качественный*), игнорируя слово *недорогой*.

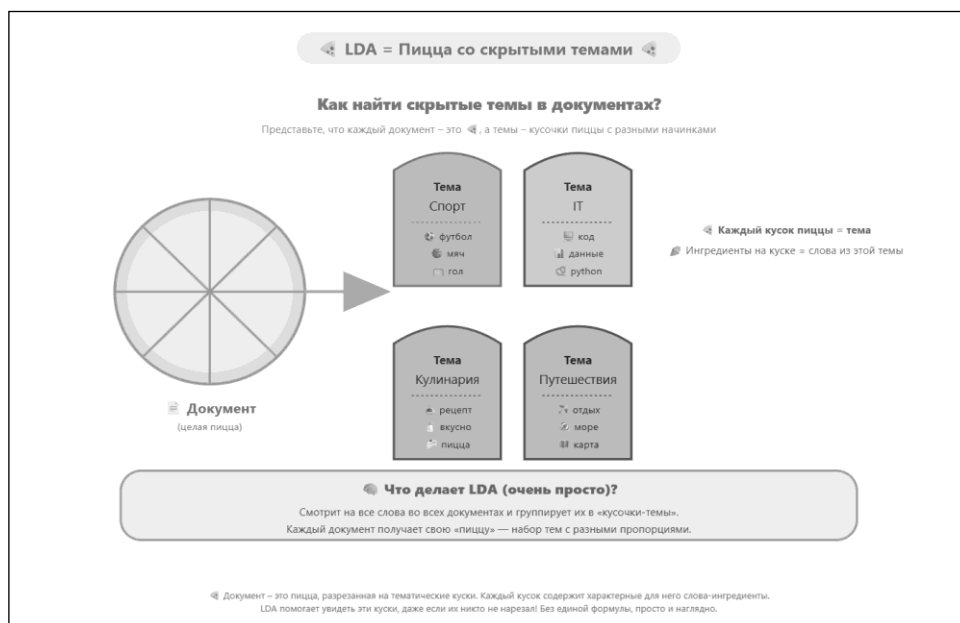
2) Скрытые марковские модели (или HMM – Hidden Markov Model) – используют для анализа последовательностей, где наблюдаемые данные (слова) являются скрытыми состояниями (например, частями речи). Она опирается на три вероятностных параметра: начальное состояние, переход между состояниями и эмиссия (наблюдение). Если разобрать предложение по частям речи (задача POS-tagging), то выражение *Я люблю программирование* модель определяет наиболее вероятную последовательность тегов: *местоимение (Я), глагол (люблю), существительное (программирование)*.

3) Байесовские модели основаны на теореме Байеса:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

В лингвистике в основном используют для классификации текстов. Пример – фильтрация спама. Вычисляется вероятность того, что письмо является спамом, при наличии определённых слов. Обучение происходит на размеченных данных. Байесовские модели просты в реализации, требуют небольшого объёма данных и хорошо работают при независимости признаков (наивное предположение). Их также применяют для определения тональности текстов, категоризации новостей и авторства.

4) Тематическое моделирование (или LDA – Latent Dirichlet Allocation) – это вероятностная модель, которая позволяет находить скрытые темы в больших коллекциях документов. Каждый документ представляется как смесь тем, а каждая тема – как смесь слов. Параметры модели оцениваются с помощью методов сэмплирования (например, Гиббса). LDA используется для кластеризации научных статей, подбора похожих новостей, анализа социальных сетей и др. В отличие от предыдущих моделей, LDA работает на уровне целых корпусов, а не отдельных последовательностей.



**Рисунок 2. – Пример LDA**

LDA в виде пиццы представлена на рисунке 2 (*документ как пицца, темы как куски*). Но нужно учесть, что такая *метафора* ломается в трёх местах.

- 1) Темы не лежат компактно. Слова разных тем перемешаны по всему тексту.
- 2) Короткие тексты получают лишние темы (из-за математических свойств LDA короткий текст часто получает несколько тем даже там, где логически одна).
- 3) Смысл слова зависит от всего корпуса. Слово *пицца* в IT-юморе может уйти в тему программирования, а не в кулинарию.

Лучше всего представлять модель LDA как *коктейль* из слов, где темы постоянно перемешиваются.

Из рассмотренных вероятностных подходов можно составить общую схему применения вероятностных моделей в компьютерной лингвистике. Выбор метода не зависит от конкретной модели. Процесс включает стандартные этапы обработки: от полученных входных данных до наиболее вероятного результата.

Более подробно представлено на рисунке 3.



**Рисунок 3. – Общая схема применения вероятностных моделей**

Вероятностные модели решают большинство задач компьютерной лингвистики. Каждая модель по-своему уникальна: n-граммы – для локального контекста, НММ – для последовательной разметки, байесовские модели – для классификации, LDA – для тематического анализа. Иногда комбинируют несколько подходов, которые являются более сложными вероятностными моделями.

Например: для извлечения именованных сущностей (NER) применяют НММ, а потом байесовский классификатор. Это делается для того, чтобы одна модель исправляла ошибки другой. НММ *грубо* находит, где начинается именованная сущность, а байесовский классификатор уточняет, что именно – имя, компания или город. Вместе они работают точнее.

#### БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Шестаков О. В. Вероятностные модели [Электронный ресурс] : конспект лекций / подгот. студентами ВМК МГУ. – Электрон. текст. дан. – М. : МГУ им. М. В. Ломоносова, [б. г.]. – 83 с. – Режим доступа: <https://teach-in.ru/file/synopsis/pdf/probabilistic-model-M.pdf>. – Дата доступа: 20.04.2026.

2. Щипицина Л.Ю. Информационные технологии в лингвистике : учеб. пособие. – М.: ФЛИНТА: Наука, 2013. – 128 с. – ISBN 978-5-9765-1431-7.

#### REVIEW OF PROBABILISTIC MODELS USED IN MODERN COMPUTATIONAL LINGUISTICS

**A.O. Varygin**

*Kuban State University*

*Krasnodar, Russia, [andrevarygin67@gmail.com](mailto:andrevarygin67@gmail.com)*

In computational linguistics, probabilistic models are used, for example, for automatic natural language processing. The choice of model is determined by the nature of the task (local context, sequential tagging, classification or corpus analysis), and their combination gives the best result.

*Keywords:* probabilistic models; computational linguistics; n-grams; hidden Markov models; Bayes' theorem; LDA.