N-грамма — последовательность из n элементов[1]. С семантической точки зрения, это может быть последовательность звуков, слогов, слов или букв. На практике чаще встречается N-грамма как ряд слов, устойчивые словосочетания называют коллокацией. Последовательность из двух последовательных элементов часто называют **биграмма**, последовательность из трёх элементов называется **триграмма**. Не менее четырёх и выше элементов обозначаются как N-грамма, N заменяется на количество последовательных элементов.

Использование N-грамм

Общее использование N-грамм

N-граммы в целом находят своё применение в широкой области наук. Они могут применяться, например, в области теоретической математики, биологии, картографии, а также в музыке. Наиболее часто использование N-грамм включает следующие области:

- извлечение данных для кластеризации серии спутниковых снимков Земли из космоса, чтобы затем решить, какие конкретные части Земли на изображении,
- поиск генетических последовательностей,
- в области генетики используются для определения того, с каких конкретных видов животных собраны образцы ДНК,
- в компьютерном сжатии,
- для индексирования данных в поисковых системах; с использованием N-грамм, как правило, индексированы данные, связанные со звуком.

Также N-граммы широко применяются в обработке естественного языка.

Использование N-грамм для нужд обработки естественного языка[править | править код]

В области обработки естественного языка N-граммы используется в основном для предугадывания на основе вероятностных моделей. N-граммная модель рассчитывает вероятность последнего слова N-граммы, если известны все предыдущие. При использовании этого подхода для моделирования языка предполагается, что появление каждого слова зависит только от предыдущих слов[2].

Другим применением N-грамм является выявление плагиата. Если разделить текст на несколько небольших фрагментов, представленных N-граммами, их легко сравнить друг с другом и таким образом получить степень сходства анализируемых документов[3]. N-граммы часто успешно используются для категоризации текста и языка. Кроме того, их можно использовать для создания функций, которые позволяют получать знания из текстовых данных. Используя N-граммы, можно эффективно найти кандидатов, чтобы заменить слова с ошибками правописания.

Пример биграммной модели

Целью построения N-граммных моделей является определение вероятности употребления заданной фразы. Эту вероятность можно задать формально как вероятность возникновения последовательности слов в неком корпусе (наборе текстов). К примеру, вероятность фразы «счастье есть удовольствие без раскаяния» можно вычислить как произведение вероятностей каждого из слов этой фразы:

```
P = P(счастье) * P(есть|счастье) * P(удовольствие|счастье есть) * P(без|счастье есть удовольствие) * P(раскаяния|счастье есть удовольствие без)
```

Чтобы определить P(счастье), нужно посчитать, сколько раз это слово встретилось в тексте, и поделить это значение на общее число слов. Рассчитать вероятность P(раскаяния|счастье есть удовольствие без) сложнее. Чтобы упростить эту задачу, примем, что вероятность слова в тексте зависит только от предыдущего слова. Тогда наша формула для расчета фразы примет следующий вид:

P = P(счастье) * P(есть|счастье) * P(удовольствие|есть) * P(без|удовольствие) * P(раскаяния|без)

Рассчитать условную вероятность P(есть|счастье) несложно. Для этого считаем количество пар 'счастье есть', и делим на количество в тексте слова 'счастье'.

В результате, если мы посчитаем все пары слов в некотором тексте, мы сможем вычислить вероятность произвольной фразы. Этот набор рассчитанных вероятностей и будет биграммной моделью.

Научно-исследовательские проекты Google[

Исследовательские центры Google использовали N-граммные модели для широкого круга исследований и разработок. К ним относятся такие проекты, как статистический перевод с одного языка на другой, распознавание речи, исправление орфографических ошибок, извлечение информации и многое другое. Для целей этих проектов были использованы текстовые корпусы, содержащие несколько триллионов слов.

Google решила создать свой учебный корпус. Проект называется Google teracorpus и он содержит 1 024 908 267 229 слов, собранных с общедоступных веб-сайтов[4].

Методы для извлечения N-грамм

В связи с частым использованием N-грамм для решения различных задач необходим надежный и быстрый алгоритм для извлечения их из текста. Подходящий инструмент для извлечения N-грамм должен быть в состоянии работать с неограниченным размером текста, работать быстро и эффективно использовать имеющиеся ресурсы. Есть несколько методов извлечения N-грамм из текста. Эти методы основаны на разных принципах:

- *Алгоритм Nagao 94* для текстов на японском[5]
- Алгоритм Лемпеля — Зива — Велча
- Суффиксный массив
- Суффиксное дерево
- Инвертированный индекс

Синтаксические N-граммы[править | править код]

Синтаксические N-граммы — это N-граммы, определяемые путями в деревьях синтаксических зависимостей или деревьях составляющих, а не линейной структурой текста[6][7]. Например, предложение: «Экономические новости оказывают незначительное влияние на финансовые рынки» может быть преобразовано в синтаксические N-граммы, следуя древовидной структуре его отношений зависимостей: новости-экономические, влияние-незначительное, влияние-на-рынки-финансовые и другие[6].

Синтаксические N-граммы отражают синтаксическую структуру в отличие от линейных N-грамм и могут использоваться в тех же приложениях, что и линейные N-граммы, в том числе в качестве признаков в векторной модели. Применение синтаксических N-грамм дает лучшие результаты при решении определенных задач, чем использование стандартных N-грамм, например, для определения авторства[8].

1. ↑ Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, pp. 54-65, February 2008. ISBN 978-80-227-2827-0.
2. ↑ *Jurafsky, D. and Martin, J.H.* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. — Pearson Prentice Hall, 2009. — 988 p. — ISBN 9780131873216.
3. ↑ Proceedings of the ITAT 2008, Information Technologies — Applications and Theory, Hrebienok, Slovakia, pp. 23-26, September 2008. ISBN 978-80-969184-8-5
4. ↑ FRANZ, Alex, BRANTS, Thorsten. Official Google Research Blog: All Our N-gram are Belong to You. Thursday, August 03, 2006 at 8/03/2006 11:26:00 AM. Созданная база N-грамм продаётся в виде 5 DVD.
5. ↑ M. Nagao and S. Mori. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), Kyoto, Japan, 1994.
6. ↑ Перейти обратно:[1] [2] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic Dependency-based N-grams as Classification Features. LNAI 7630, pp. 1-11, 2012.
7. ↑ Grigori Sidorov. Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction. International Journal of Computational Linguistics and Applications, Vol. 4, No. 2, pp. 169—188, 2013.
8. ↑ Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic N-grams as Machine Learning Features for Natural Language Processing. Expert Systems with Applications, Vol. 41, No. 3, pp. 853—860, DOI 10.1016/j.eswa.2013.08.015.