Обработка естественного языка

Прежде, чем перейти к трансформерам, сделаем быстрый обзор того, что такое обработка естественного языка (NLP), и почему мы заинтересованы в этой сфере.

Что такое NLP?

NLP - область лингвистики и машинного обучения, которая изучает все, что связано с естественными языками. Главная цель NLP не просто понимать отдельные слова, но и иметь возможность понимать контекст, в котором эти слова находятся.

Список типичных NLP-задач с некоторыми примерами:

- **Классификация предложений**: определить эмоциональную окраску отзыва, детектировать среди входящих писем спам, определить грамматическую правильность предложения или даже проверить, являются ли два предложения связанными между собой логически
- **Классификация каждого слова в предложении**: вычленить грамматические составляющие предложения (существительное, глагол, прилагательное) или определить именованные сущности (персона, локация, организация)
- **Генерация текста**: закончить предложение на основе некоторого запроса, заполнить пропуски в тексте, содержащем замаскированные слова
- **Сформулировать ответ на вопрос**: получить ответ на заданный по тексту вопрос
- **Сгенерировать новое предложение исходя из предложенного**: перевести текст с одного языка на другой, выполнить автоматическое реферирование текста

NLP не ограничивается только письменным текстом. Есть множество сложных задач, связанных с распознаванием речи и компьютерным зрением, таких как транскрибирование аудио или описание изображений.

Почему это сложно?

Компьютеры не обрабатывают информацию так же, как люди. Например, когда мы читаем предложение «Я голоден», мы можем легко понять его значение. Точно так же, имея два предложения, такие как «Я голоден» и «Мне грустно», мы можем легко определить, насколько они похожи. Для моделей машинного обучения (ML) такие задачи сложнее. Текст должен быть обработан так, чтобы модель могла учиться на нем. А поскольку язык сложен, нам нужно тщательно продумать, как должна выполняться эта обработка. Было проведено много исследований того, как представлять текст, и мы рассмотрим некоторые методы в следующей главе.