**SMT (Statistical Machine Translation)**

SMT основывается на байесовской теореме для перевода текста с одного языка на другой. Основная идея заключается в том, чтобы максимизировать вероятность правильного перевода исходного предложения на целевой язык.

**Основные компоненты:**

1. **Корпус параллельных текстов –** набор предложений на двух языках, где каждое предложение на одном языке соответствует переводу этого предложения на другой язык
2. **Языковая модель (Language Model) –** модель, которая оценивает вероятность последовательности слов на целевом языке. Она помогает определить, насколько естественно звучит предложенный перевод.
3. **Модель выравнивания (Alignment Model) –** определяет соответствие между словами и фразами в исходном тексте и их переводами в целевом тексте. Выравнивание может быть однословным (каждое слово соответствует одному слову) или фразовым (группы слов соответствуют друг другу).
4. **Модель перевода (Translation Model) –** оценивает вероятность того, что конкретная последовательность слов на исходном языке будет переведена в определенную последовательность слов на целевом языке.

**Важная терминология:**

* **Байесовская теорема –** используется для вычисления вероятности того, что данное предложение на исходном языке будет переведено в определенное предложение на целевом языке.
* **EM-алгоритм (Expectation-Maximization Algorithm) –** алгоритм, используемый для оптимизации параметров моделей SMT. Он состоит из двух этапов: этапа ожидания (expectation) и этапа максимизации (maximization).
* **Токенизация –** процесс разделения текста на отдельные слова или токены.
* **Векторизация –** преобразование текстовых данных в числовые представления, которые могут быть использованы алгоритмами машинного обучения.
* **Корпус –** набор текстов, используемых для обучения модели.
* **Параллельные корпуса – к**орпуса, содержащие тексты на разных языках, где каждый текст имеет соответствующий перевод на другой язык.
* **Оценочная функция –** функция, которая оценивает качество перевода, сравнивая его с эталонными переводами.

**Пример использования:**

1. **Подготовка данных:**
   * Загружаются параллельные корпуса текстов.
   * Тексты токенизируются и векторизуются.
2. **Обучение модели:**
   * Языковая модель обучается на целевых текстах.
   * Модель выравнивания обучается на параллельных корпусах.
   * Модель перевода обучается на основе оценочной функции.
3. **Перевод:**
   * Исходный текст подается в модель.
   * Модель генерирует возможные переводы и выбирает наиболее вероятный вариант.