**Принцип работы SMT (основные компоненты, терминология)**

**SMT (Statistical Machine Translation)** - это метод машинного перевода, основанный на статистическом анализе параллельных текстов.

математика: дано х, найти более правдоподобный у

Основные компоненты SMT:

1. **Параллельные корпуса** - набор текстов на разных языках, соответствующие по смыслу.

2. **Модель перевода (t-model)** - статистическая модель, которая определяет вероятность перевода слова или фразы с одного языка на другой. У может быть искаженным, ищем наиболее правдоподобные соотвествия.

3. **Языковая модель** - статистическая модель, которая оценивает вероятность последовательности слов в целевом языке.

4. **Декодер** - алгоритм, который собирает наиболее грамматичные и лексически правдоподобные результаты.

Терминология:

* Источник (source) - исходный текст, который нужно перевести.
* Цель (target) - целевой текст, на который нужно перевести исходный текст.
* **Выравнивание (alignment)** - соответствие между словами или фразами в исходном и целевом текстах.

Используемые библиотеки:

* Библиотеки для обработки текстов: NLTK, Spacy (можно провести пословную токенизацию каждой фразы, словарь уникальных словофом)
* Библиотеки для машинного обучения: Scikit-learn
* Оценка результатов: NLTK Translate

Принцип работы SMT заключается в анализе статистических связей между словами и фразами на разных языках для построения наиболее вероятного перевода.

почитать: Учебник "Statistical Machine Translation", Philipp Koehn

с курса по комплингв: <https://colab.research.google.com/drive/1ilLmABg24bSeT_woHZ6M3u9CsEUynyXT#scrollTo=PNcqDYfAg0oF>