Статистический машинный перевод (англ. Statistical machine translation — SMT) — разновидность машинного перевода, где перевод генерируется на основе статистических моделей, параметры которых являются производными от анализа двуязычных корпусов текста (text corpora).

Наиболее правдоподобный перевод, наиболее вероятное соответствие фразы в параллельных корпусах

SMT:

1. ***N-граммная языковая модель*** - моделируем *вероятностное распределение* конструкций на уровне слов или фраз в языке Y
2. ***Модель перевода (t-model)*** - собираем *статистику* соотвествий фраз в паралелльном корпусе, ищем переводческие соответствия X - Y и моделируем их с помощью теории *вероятности*:
3. Допускаем, что любое предложение языка Y может быть "искаженной" версией некой фразы на языке X
4. Ищем наиболее правдоподобные соответствия X - Y
5. ***Декодер*** - ищем наиболее грамматичные и лексически правдоподобные результаты, отбираем среди *гипотез* один результат

Препроцессинг (данные)

- OPUS Corpora
- Kaggle
- Hugging Face

Подготовка данных:
Токенизация, словарь уникальных словоформ