**SMT**

(*Statistical machine translation)*

**Краткий** смысл статистического перевода: если предложение со словом Х часто сопоставляется предложению со словом Y, компьютер предположит, что Х=Y.

“Параллельные корпуса” - реально работающая вещь, которая облегчает работу (нам - лингвистам).

Статистический машинный перевод строится на “параллельных корпусах” и “вероятности”.

Нам доступен некий параллельный корпус, на основе которого мы:

1. делаем обучающую выборку
2. предобработку текста
3. создаем словарь уникальных словоформ
4. занимаемся уже непосредственно подсчетами вероятности
5. t-model (статистика соответствий фраз в параллельном корпусе - поиск переводческих соответствий Х-У, основываясь на вероятности!)
6. декодер (среди всех гипотез выбираем лучшее)
7. оценка результатов

Родоначальник SMT - Уоррен Уивер. Первые попытки были в 1949 г., в числе прочего он применял информационную теорию Клода Шэннона[[1]](#footnote-0). 1991 г. - снова интересно, компания IBM.

На сегодняшний день это самый изучаемый метод машинного перевода.

* SMP :
* SMP не привязаны к определенной языковой паре.
* SMP на основе правил требуют ручной разработки лингвистических правил, и при этом такие правила привязаны к одному языку.
* Итоговые тексты, полученные SMP, наиболее близки к переводам человека.

1. В 1949 году Шеннон вместе с Уорреном Уивером разработал математическую модель коммуникации, описывающую линейную передачу посланий. Источник информации должен иметь передатчик, а получатель — приемник. Между передатчиком и приемником идет сигнал. Шеннон ввел понятие информационного шума, искажающего информацию в ходе передачи, использовав для его характеристики физически-философскую категорию «энтропия». Схемы, предложенные Шенноном и Уивером, легли в основу информатики и кибернетики. [↑](#footnote-ref-0)