Конспект SMT (Statistical Machine Translation)

Ключевые персоны:

1. Параллельные корпуса – наборы текстов на двух языках, где каждое предложение на одном языке соответствует предложению-переводу на другом.

2. Языковая модель (LM) – оценивает вероятность последовательности слов на ***целевом языке*** (Y). Использует ***n-граммы*** для предсказания вероятности следующего слова.

3. Модель перевода (Translation Model, t-model) – собирает статистику соответствий между языками X и Y, предполагая, что предложение на Y может быть "искаженной" версией фразы на X.

4. Декодер – алгоритм, выбирает наиболее грамматически и лексически правдоподобный перевод из множества гипотез.

Сюжет

Дано предложение на языке X, нужно найти наиболее правдоподобный перевод на язык Y: \ [\text{argmax } P(y|x) = \text{argmax } P(y) \cdot P(x|y) \]

- \(P(y)\) – вероятность предложения на языке Y (языковая модель).

- \(P(x|y)\) – вероятность того, что предложение на X является переводом предложения на Y (модель перевода).

1. Сбор данных: параллельные корпуса текстов на двух языках.

2. Токенизация и векторизация: разбиение текста на слова/токены и преобразование в числовые данные.

3. Обучение моделей:

- Языковая модель обучается на текстах языка Y.

- Модель перевода обучается на параллельных корпусах.

4. Перевод:

- Исходный текст подается в модель.

Акты (содержание пьесы)

1. Сбор параллельных корпусов – тексты на двух языках, выровненные по предложениям.

2. Создание языковой модели – вычисление вероятностей последовательностей слов на языке Y.

3. Сопоставление фраз – анализ параллельного корпуса для определения вероятностей \(P(x|y)\).

4. Максимизация произведения \(P(y) \cdot P(x|y)\) – поиск наиболее вероятного перевода.

5. Декодирование – выбор лучшего перевода из возможных вариантов.

Герои:

- N-грамма – последовательность из N слов, используемая для оценки вероятности появления слов в языковой модели.

- Translation Probability – вероятность того, что фраза на Y является переводом фразы на X.

- Alignment – соответствие слов или фраз между исходным и целевым текстами.

- Decoding – процесс поиска лучшего перевода.

- BLEU (Bilingual Evaluation Understudy) – метрика оценки качества перевода, сравнивающая машинный перевод с эталонным.

- EM-алгоритм (Expectation-Maximization) – метод оптимизации параметров моделей SMT:

- E-шаг (Ожидание): вычисление скрытых данных.

- M-шаг (Максимизация): обновление параметров модели.

Библиотеки:

- Scikit-learn – для машинного обучения.

- nltk – для обработки текста.

- pandas – для работы с табличными данными (например, очистка корпусов).