

## 18. Практическое занятие: Алгоритмы машинного перевода и нейронные системы

### Цель занятия

Изучить эволюцию алгоритмов машинного перевода (от правил до нейросетей), понять архитектуру систем NMT (Neural Machine Translation) и научиться оценивать качество перевода для разных языковых пар.

### Теоретический минимум

1. **RBMT (Rule-Based):** Перевод на основе лингвистических правил и словарей. Точный в грамматике, но «деревянный» в стиле.
2. **SMT (Statistical):** Статистический перевод. Основан на вероятностях того, что фраза \$A\$ соответствует фразе \$B\$ в параллельном корпусе.
3. **NMT (Neural):** Нейронный перевод. Использует архитектуру Encoder-Decoder. Текст переводится в абстрактное векторное пространство (context vector) и генерируется на другом языке.
4. **Zero-Shot Translation:** Способность нейросети переводить между языковыми парами, для которых не было прямых параллельных данных при обучении.

### Задание 1. Сравнение языковых пар (Морфологическая дистанция)

Качество перевода сильно зависит от того, насколько «похожи» языки в паре.

- **Близкие языки:** Русский — Украинский, Английский — Немецкий.
- **Дистантные языки:** Английский — Узбекский, Русский — Японский.

### Инструкция:

Проанализируйте фразу «Я иду в школу» в двух парах:

1. **En-De (SVO-структура):** I go to school -> Ich gehe zur Schule.
2. **En-Uz (Агглютинация):** I go to school -> Men maktabga boraman (направление движения -ga приклеивается к слову).

**Вопрос:** Почему нейросетям сложнее переводить на агглютинативные языки (узбекский, турецкий)?

### Задание 2. Работа с Encoder-Decoder архитектурой (NMT)

В нейронном переводе Encoder сжимает предложение в фиксированный вектор, а Decoder разворачивает его в целевой язык.

### Инструкция:

Используйте библиотеку transformers для выполнения перевода с помощью модели Helsinki-NLP (популярные легковесные модели для разных пар).

## Python

```
from transformers import MarianMTModel, MarianTokenizer

def translate_text(text, model_name="Helsinki-NLP/opus-mt-en-ru"):
    tokenizer = MarianTokenizer.from_pretrained(model_name)
    model = MarianMTModel.from_pretrained(model_name)

    # Подготовка текста
    inputs = tokenizer(text, return_tensors="pt", padding=True)
    # Генерация перевода
    output = model.generate(**inputs)
    # Декодирование
    return tokenizer.decode(output[0], skip_special_tokens=True)

print("Перевод (EN-RU):", translate_text("Machine learning is a subset of artificial intelligence."))
```

## Задание 3. Оценка качества: Метрика BLEU

Метрика BLEU (Bilingual Evaluation Understudy) измеряет точность перевода, сравнивая количество совпадающих n-грамм в машинном переводе и эталоне.

$$BLEU = BP \cdot \exp \left( \sum_{n=1}^N w_n \ln p_n \right)$$

Инструкция:

Рассчитайте точность для биграмм (2-gram):

- **Эталон:** «Кошка сидит на ковре»
- **Перевод:** «Кошка сидит под ковром»
- **Общие биграммы:** («Кошка сидит», «сидит под» — нет, «под ковром» — нет).
- **Результат:** Только 1 биграмма совпала из 3 возможных.

## Задание 4. Исследование проблемы «Gallicisms» и галлюцинаций

Нейросети иногда «галлюцинируют», добавляя факты, которых нет в оригинале, или путая род слов.

Инструкция:

Попробуйте перевести предложение с гендерной двусмысленностью с английского на русский: "The doctor asked the nurse to help him".

1. Как модель определила пол врача и медсестры?
2. Насколько это зависит от статистических смещений (bias) в обучающих данных?

#### Задание 5. Анализ парных языков (Low-Resource Languages)

Существуют языки с огромными параллельными корпусами (En, Es, Fr) и «малоресурсные» языки (многие африканские и центральноазиатские диалекты).

Инструкция:

Объясните концепцию Back-translation: как можно использовать мооязычные данные на узбекском языке, чтобы улучшить перевод с английского на узбекский?

#### Контрольные вопросы

1. В чем главное преимущество NMT перед SMT (статистическим переводом)?
2. Что такое «Bottle-neck» (узкое горлышко) в архитектуре Encoder-Decoder и как его решил механизм Attention?
3. Почему профессиональные переводчики до сих пор используют CAT-инструменты (Computer-Assisted Translation), а не просто копируют текст из нейросетей?

Итог работы

Вы изучили внутреннее устройство современных систем перевода. Эти знания позволяют не только использовать API переводчиков, но и понимать ограничения моделей при работе со сложными языковыми парами, а также настраивать собственные системы для узких доменных областей (медицина, право).