Обучение машинного перевода без параллельных текстов*

Ямалутдинов А. В., Стрижов В. В., Бахтеев О. Ю.
yamalutdinov.av@phystech.ru, strijov@ccas.ru, bakhteev@phystech.edu
[1] Московский физико-технический институт, Москва, Россия

В данной работе исследуется метод обучения без учителя для задачи машинного перевода. Рассматривается подход, основанный на автокодировщиках: каждое предложение отображается кодировщиком в вектор в латентном пространстве, а декодировщик восстанавливает полученный вектор в предложение на другом языке. Оптимизация моделей проводится таким образом, чтобы скрытые пространства автокодировщиков для разных языков совпадали. В качестве исходного представления предложений предлагается рассматривать их графовое описание, получаемое с использованием мультиязычных онтологий.

Ключевые слова: *машинный перевод, автокодировщики, нейронные сети.*

1 Введение

В последнее время наблюдаются значительные успехи в решении задачи машинного перевода, главным образом за счет использования глубоких нейронных сетей. Основным недостатком такого подходя является тот факт, что для обучения таких сетей требуется огромное (около миллиона) количество параллельных предложений. Это требование значительно ограничивает возможности построения моделей для низкоресурсных языков (т.е. языков, для которых данных в открытом доступе немного).

Исследуется возможность решения задачи машинного перевода как задачи обучения без учителя. В таком случае нет необходимости в большом количестве параллельных предложений, для обучения модели достаточно иметь выборку предложений на каждом из языков.

*Задачу поставил: Стрижов В. В. Консультант: Бахтеев О. Ю.