Обучение машинного перевода без параллельных текстов.*

Саттаров Т. И.

sattarov.ti@phystech.edu

[1]Московский Физико-технический институт

Рассматривается метод обучения без учителя для задачи машинного перевода. В данном подходе используется нейросетевая модель с применением seq2seq. Каждое предложение на рассматриваемых языках отображается кодировщиком в вектор из скрытого пространства, а декодировщик восстанавливает предложение из вектора скрытого пространства. Для дальнейшего улучшения модели перевода предлагается применить модель T2T, позволяющую получить многоуровневые зависимости между словами в исходных предложениях. Качество работы проверяется на парах языков «английский-казахский» и «русский-украинский».

Ключевые слова: *Машинный перевод, seq2seq, T2T, обучение без учителя.*

1 Введение

Для решения задачи автоматического перевода с одного языка на другой существует несколько подходов: перевод по правилам, статистический перевод и др. Статистическая модель разделяет переводимое предложение на отдельные слова и фразы, перебирает все варианты перевода для каждого фрагмента и взвешивает вероятность каждого из них, исходя из того, какой вариант встречался чаще на обучающей выборке. Этот метод хорош для редких фраз и коротких предложений, так как он «не видит» всё длинное предложение в целом. Благодаря недавним достижением в области глубинного обучения и доступности большого количества параллельных текстов машинный перевод стал очень эффективным на некоторых парах языков. Не смотря хорошую точность перевода у этого метода есть большой недостаток: необходимость большого числа параллельных текстов — количество предложений исчисляется миллионами. К сожалению, построение параллельных текстов

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Бахтеев О. Ю. Консультант: Бахтеев О. Ю.

в таком количестве требует большого количества ресурсов, если таковых данных нет в открытом доступе.

В нашей статье мы исследуем возможность обучения машинного перевода без учителя. Единственное, что мы требуем от данных, чтобы у нас было достаточное количество одноязычной информации. Основная идея состоит в том чтобы построить общее для обоих языков «скрытое» пространство удовлетворяющее следующим свойствам. Единственное, что мы требуем от данных, чтобы у нас было достаточное количество одноязычной информации. Основная идея состоит в том чтобы построить общее для обоих языков «скрытое» пространство удовлетворяющее следующим свойствам: во-первых, модель должна уметь восстанавливать предложение в исходном языке исходя из зашумленной версии его в «скрытом» пространстве, во-вторых, , модель учится восстанавливать исходное предложение по его зашумленному переводу. Все же использование seq2seq не позволяет хорошо работать с окончаниями, так как, например, слова в разных падежах русского языка для нее различны, что приводит к разрозненности данных. Чтобы исправить это предлагается использовать T2T модель, позволяющую получить многоуровневые зависимости между словами в исходных предложениях.

Литература