А. Калацкая

ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В АВТОМАТИЧЕСКОМ ПЕРЕВОДНОМ СЛОВАРЕ

Нейронный машинный перевод – новый подход, который адресует проблемы, существующие в рамках статистического машинного перевода: игнорирование зависимостей, которые находятся на большом расстоянии друг от друга, и комплексность, так как добавляется все больше и больше характеристик для улучшения работы систем статистического машинного перевода. Таким образом, нейронный машинный перевод направлен на решение проблем машинного перевода с использованием нейронных сетей.

Нами был задействован англо-русский параллельный корпус текстов: United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. Данный корпус состоит из 2100 резолюций Генеральной Ассамблеи ООН с переводом на шесть официальных языков Организации Объединенных Наций, из которых для проведения исследования была извлечена англо-русская языковая пара. В корпусе каждое английское предложение представлено с новой строки и имеет соответствующее ему предложение-перевод на русский язык. В рамках работы с данными были выбраны пары предложений, которые меньше или равны 30 словам.

Весь корпус текстов был разделен на тренировочную выборку, оценочную выборку (100 предложений) для подсчета качества перевода с текущими весами на промежуточном этапе и тестовая выборка для итоговой оценки перевода. В тренировочной выборке из всех исходных и целевых предложений формируются словари уникальных слов: словарь уникальных слов целевого русского текста, словарь уникальных слов исходного англоязычного текста, словарь всех слов.

Тренировочная выборка, состоящая из англо-русских пар предложений, будет являться основой для обучения данной модели нейронного машинного перевода. Для обработки и последующего представления лингвистической информации, в нашем случае слов, будет использован слой нейронной сети, который обеспечит векторное представление входных и выходных значений. То есть каждое слово будет встроено в пространство заданной размерности и представлено в виде вектора в нем. Вложения – векторные представления каждого элемента – будут натренированы как параметры функции внутри нейронной сети и в процессе обучения будут менять координаты в пространстве, основываясь на встречаемости соответствующих слов в контексте.

Таким образом, для обработки лингвистических единиц был использован слой нейронной сети, который определил векторное представление входных и выходных значений и сформировал две модели языка: английскую языковую модель и русскую языковую модель, основываясь на контекстной встречаемости в корпусе.