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 работе рассматривается метод обучения без учителя для задачи построения системы машинного перевода. Рассматривается подход, основанный на автокодировщиках: каждое предложение переводится кодировщиком в вектор таким образом, чтобы скрытые пространства кодировщиков для разных языков совпадали. Далее, декодировщик переводит полученное векторное представление в предложение на другом языке. Предлагается модификация данного подхода с использованием информации, полученной от систем мультиязычных онтологий: для каждого переводимого предложения строится граф, ребра которого соответствуют отношениям между словами внутри онтологии. Предлагается дополнительный регуляризатора оптимизации —- функция, штрафующая модель перевода за несоответствие графового представления исходного и переведенного предложения. Для анализа предложенной модификации проводится эксперимент на языковой паре "английский-французский".

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

1 Введение

В настоящее время одним из основных подходов к задаче машинного перевода является использование глубоких нейронных сетей. Традиционный подход к решению данной задачи предполагает, что модель обучается на корпусе параллельных текстов (обучающая выборка состоит из пар предложений на разных языках) [?]. Однако в таких моделях для достижения высокого качества перевода необходимы корпуса, состоящие из, как правило, нескольких миллионов параллельных предложений [?].

Для некоторых пар языков не существует обучающей выборки достаточного размера. Для решения проблемы машиного перевода между данными языками было предложенно несколько подходов, в частности, подход, основанный на автокодировщиках. [?] [?] Рассматриваются автокодировщики, реализованные в виде рекуррентных нейронный сетей, который оптимизируются таким образом, чтобы скрытые пространства автокодировщиков, кодирующих текст на разных языказ, совпадали.

В данной статье предлагается усовершенствовать предложенный метод, используя данные о мультиязычных онтологиях [?]. Для исходного и переведенного предложения предлагается строить графовые представления, ребра которых соответствуют отношениям между словами. Далее эти представления планируется использовать для регуляризации модели, т.е. роста функционала ошибки в случае, когда графовые представления исходного и переведенного предложения значительно отличаются.

В качестве эксперимента планируется перевод предложений между парой языков "английский-французский".

2 Постановка задачи