В. Лапша

ИННОВАЦИОННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПИСЬМЕННОМ ПЕРЕВОДЕ

Информационные технологии играют все более важную роль в современном мире. Во многом это касается и перевода, в котором за последние 10 лет появились новые технологии и методы, позволившие значительно улучшить качество конечного продукта. Одним из самых популярных методов, используемых во многих электронных системах перевода (в т.ч. в системе перевода Google), является *векторное представление слов* (или *word embedding*), основанное на дистрибутивной семантике, которое представляет собой общее название для различных подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам из некоторого словаря векторов небольшой размерности.

Существует ряд методов моделирования языка, эффективно применяющихся в областях, где количество текстов мало, а словарь ограничен. Однако с развитием современных технологий возникла потребность в более совершенных методах. Один из них – метод «word2vec», предложенный Т. Миколовым в 2013 году. Принцип работы заключается в следующем: «word2vec» принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала программа создает словарь, «обучаясь» на входных текстовых данных, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (следовательно, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.

Метод позволяет установить семантическую близость между словами, тем самым научить компьютер «понимать» контекст. Используя данный метод, можно относительно просто избежать перевода одного слова разными способами, что особенно полезно в случаях, когда несколько переводчиков делят перевод крупного текста между собой, а затем соединяют получившиеся части в цельный перевод.

Разумеется, эффективность и точность векторных моделей целиком зависит от выборки: если обучить программу на серии рассказов К. Льюиса, то слова *девочка* и *лев* будут ближе друг к другу, чем слова *девочка* и *кукла*. Модель при этом будет работать корректно, но не будет соответствовать действительности. К сожалению, подобные ситуации встречаются и при использовании для обучения менее специфичных корпусов. Например, «word2vec», обученная на текстах Google News, сильно подвержена стереотипам: слово *почетный* ближе к слову *мужчина*, а *подчиняющаяся* – к слову *женщина*. Для минимизации подобных казусов программисты рекомендуют расширять выборку текстов и стараться избегать чрезвычайных случаев.