Коррекция векторных представлений слов для улучшения их семантической близости

А.М. Колосов[1], А.И. Майсурадзе[1]

[1]*Московский государственный университет им. М.В. Ломоносова, Ленинские горы, 1, стр. 52, Москва, Россия, 119991*

Аннотация

Трудоёмкость методов получения векторных представлений слов постоянно возрастает, что связано как с увеличением объёма входных данных, так и с усложнением моделей. Разработка и обучение новой модели, которая бы учитывала дополнительную экспертную информацию о семантической близости между словами, сегодня оценивается в тысячи часов вычислительных экспериментов. В данном исследовании существующие представления корректируются путём их преобразования во вторичные представления с учетом экспертной информации о семантической близости. Предложена функция преобразования представлений и метод её обучения. Таким образом, вместо построения новой модели проводится коррекция существующей, что существенно снижает вычислительные затраты по сравнению с разработкой новой модели.

Ключевые слова

Векторные представления, семантическая близость, реализация расстояний

1. Введение

Если исторически первые модели представления слов не предполагали обучения, то сегодня в подавляющем большинстве случаев настройка модели предполагает оптимизацию некоторого функционала качества на некоторой обучающей информации. Если возникает обучающая информация нового типа, то необходимо менять если не саму архитектуру модели, то по крайней мере функционал качества и метод её обучения. Таким образом, полный цикл разработки и настройки модели представления слов, учитывающей дополнительный тип обучающей информации, является очень трудоемким. Например, авторы fasttext [2] оценивают трудоемкость создания его следующей версии в миллионы часов вычислений.

В данной работе рассматривается компромиссный подход, когда вместо создания принципиально новой модели векторного представления слов мы корректируем результаты работы существующей модели с учетом дополнительной обучающей информации. В качестве такой дополнительной информации используются экспертные оценки семантической близости между словами [3]. Научная ценность данной работы состоит не только в построении конкретной корректирующей операции, но и в предварительной выработке функционалов качества и методов обучения для будущих более богатых моделей, которые сразу будут использовать обучающую информация разного типа.

Отметим, что сегодня распространены как модели, позволяющие получить представления для изолированных слов [4], так и модели, дающие представление слова с учетом конкретного контекста [5]. В данном исследовании могут быть актуальны оба типа моделей: для моделей, позволяющих получать контекстно-зависимые представления слов, в функционале качества мы используем бесконтекстные представления слов.

2. Улучшение векторных представлений

Экспертная информация о семантической близости слов обладает рядом особенностей. Во-первых, она задана на сравнительно малом числе слов, при этом далеко не все возможные комбинации слов образуют пары. Во-вторых, шкала оценок слов не является арифметической: смысл имеет порядок оценок, а не их абсолютное значение. Указанные особенности требуют учета при разработке функционала качества и метода обучения.

Предложенный метод в чем-то аналогичен известному подходу triplet-loss. Особенности состоят в формировании групп объектов и вычислении на них функции потерь. Функционал качества — это среднее всех потерь по всем сформированным группам объектов. Обучается преобразование первичных векторных представлений во вторичные. Вид этого преобразования позволяет соблюдать баланс между первичными представлениями и дополнительными требованиями.

Отбор групп объектов определяется разреженностью пар слов, для которых доступна экспертная информация. Функция потерь опирается на порядок экспертных оценок, но не их абсолютную величину, что соответствует задачам реализации расстояний [1].

3. Заключение

В работе описан подход к улучшению семантической близости векторных представлений слов за счёт использования внешней информации. Экспертная информация о семантической близости сегодня используется только для проверки качества уже построенных векторных представлений слов, но не для самого построения таких представлений. В данном исследовании экспертная информация о семантической близости использована для коррекции векторных представлений, которые первично были построены без нее.

4. Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 20-01-00664-а) и госбюджетной темы НИР № 5.1.21 МГУ имени М. В. Ломоносова.

5. Литература

[1] Chung, F. Distance realization problems with applications to Internet tomography / F. Chung, M. Garrett, R. Graham, D. Shallcross // Journal of Computer and System Sciences. – 2001. – Vol. 63(3). – P. 432-448.

[2] Bojanowski, P. Enriching Word Vectors with Subword Information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov. – 2016.

[3] Agirre, E. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches / E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa // Proceedings of NAACL-HLT. – 2009.

[4] Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // Proceedings of Workshop at ICLR. – 2013.

[5] Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. – 2018.