Улучшение качества векторных представлений слов за счёт использования нескольких источников представлений

А.М. Колосов
Московский государственный университет им. М. В. Ломоносова, факультет ВМК
Москва, Россия
akolosov@cs.msu.ru

А.И. Майсурадзе
Московский государственный университет им. М. В. Ломоносова, факультет ВМК
Москва, Россия
maysuradze@cs.msu.ru

Векторные представления слов активно используются в задачах машинного перевода, рекомендательных системах и информационном поиске. В данном исследовании проверяется гипотеза о том, что в четвёрках слов, для которых несколькими независимыми методами были получены одинаковые порядки на расстояниях между словами, монотонных четвёрках, содержится информация об истинном порядке для четвёрок с разным порядком, антимонотонных четвёрок. Проверяется, что в случае определения истинного порядка и построения векторных представлений на основе исходных и восстановленных монотонных четвёрок, качество векторных представлений слов повышается. Предложены метод отбора четвёрок слов, модель построения скорректированных векторных представлений слов и способ сравнения качества исходных и полученных в ходе коррекции векторных представлений слов.

Векторные представления, семантическая близость, слияние данных

1. Введение

Качество векторных представлений слов напрямую влияет на эффективность решения задач машинного перевода, работу рекомендательных систем и систем информационного поиска. За прошедшее десятилетие появились и развились несколько семейств методов построения векторных представлений слов, позволяющих получать представления всё более высокого качества. Вначале появилось семейство вычислительно эффективных методов Skipgram и CBOW [1], потом FastText [2], позволяющее получать представления слов при расширении словаря, и затем Bert [3], позволяющее получать представления для последовательностей слов с учётом контекста. В результате, трудоёмкость построения векторных представлений слов возрастает, однако их качество, оцениваемое как ранговая корреляция с экспертными оценками семантической близости между словами [4], не превышает 0.8 [5].

В работе рассматривается подход, когда вместо разработки нового метода построения векторных преставлений слов, проводится слияние нескольких исходных, первичных представлений слов, в одни вторичные представления, что с одной стороны существенно снижает затраты по сравнению с разработкой нового метода построения первичных представлений, а с другой позволяет получить представления слов более высокого качества. Научная ценность работы состоит не только в предложенном подходе к получению вторичных представлений из нескольких наборов первичных, но и в самих полученных вторичных представлениях слов более высокого качества, поскольку размер словаря имеет тот же порядок близости по количеству слов, что и естественный язык.

2. Слияние векторных представлений

Предложенный подход использования нескольких наборов вторичных представлений близок известному подходу triplet-loss. Также формируются группы объектов и на них вычисляется функция потерь. Функционал качества — это среднее всех потерь по всем сформированным группам объектов. Обучается преобразование объединённых первичных представлений во вторичные. Функция потерь опирается на порядок близостей в парах объектов, но не их абсолютную величину. Отбор групп объектов производится за счёт сэмплирования так, чтобы определить группы, в которых совпадают порядки близостей в парах объектов в различных группах.

Полученные экспериментальные результаты подтверждают, что качество векторных представлений слов, определяемое как ранговая корреляция с экспертными наборами данных, представленными в средстве GluonNLP [6], в результате слияния нескольких первичных представлений в одни вторичные, может увеличиваться.

3. Заключение

В работе описан подход к улучшению качества векторных представлений слов за счёт использования нескольких источников представлений. В данном исследовании первичные векторные представления используются как источник сведений о близостях в парах слов для построения вторичных векторных представлений более высокого качества. Описанный подход получения векторных представлений слов более высокого качества может быть обобщён для улучшения качества векторных представлений объектов других модальностей.

Благодарности

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

Литература

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

[2] Bojanowski, P. Enriching Word Vectors with Subword Information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // Transactions of the Association for Computational Linguistics. – 2017. –Vol. 5. – P. 135–146

[3] Devlin, J. / BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, L. Kenton, K. Toutanova // Proceedings of the 2019 Conference of the North American. – 2018. – Vol. 1. – P. 4171–4186.

[4] 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 // In Proceedings of NAACL-HLT. – 2009. – P. 19-27.

[5] Word Embedding [Electronic resource]. — Access mode: https://nlp.gluon.ai/model_zoo/word_embeddings/index.html (22.11.2022)

[6] Word Embedding Evaluation Datasets [Electronic resource]. — Access mode: https://nlp.gluon.ai/api/data.html#word-embedding-evaluation-datasets (22.11.2022)