DOI 10.37882/2223–2966.2022.04.39

ОБУЧЕНИЕ СЛОВАРЮ С ПОМОЩЬЮ ОПТИМАЛЬНОГО ТРАНСПОРТА ДЛЯ НЕЙРОННОГО МАШИННОГО ПЕРЕВОДА

VOCABULARY LEARNING VIA OPTIMAL TRANSPORT FOR NEURAL MACHINE TRANSLATION

Zhong Ruiyu

Summary. The choice of token vocabulary affects the performance of machine translation. This paper aims to figure out what is a good vocabulary and whether one can find the optimal vocabulary without trial training. To answer these questions, we first provide an alternative understanding of the role of vocabulary from the perspective of information theory. Motivated by this, we formulate the quest of vocabularization — finding the best token dictionary with a proper size — as an optimal transport (OT) problem. We propose VNMT, a simple and efficient solution without trial training. Empirical results show that VNMT outperforms widely-used vocabularies in diverse scenarios, including WMT-14 English-German and TED multilingual translation. For example, VNMT achieves almost 70% vocabulary size reduction and 0.5 BLEU gain on English-German translation.

Keywords: natural language processing, machine translation, vocabulary, optimal transport, machine learning, multilingual translation.

Чжун Жуйюй
Аспирант, ФГАОУ ВО «Национальный исследовательский университет «Московский институт электронной техники»
zry1988510@126.com

Аннотация. Выбор основных признаков словаря влияет на производительность машинного перевода. Эта работа стремится выяснить, что такое хороший словарный запас и можно ли найти оптимальный словарный запас без пробного обучения. Чтобы ответить на эти вопросы, мы сначала обеспечиваем альтернативное понимание роли словарного запаса с точки зрения теории информации. Исходя из этих предпосылок мы будем искать словарь с наилучшими признаками подходящего размера — как оптимальный транспорт (OT). Мы предлагаем VNMT, простое и эффективное решение без пробного обучения. Эмпирический результаты показывают, что VNMT превосходит широко используемые словари в различных сценариях, включая WMT-14 английский-немецкий и TED многоязычный перевод. Например, VNMT достигает почти 70% сокращения размера словарного запаса и 0,5 BLEU усиления в англо-немецком переводе.

Ключевые слова: обработки естественного языка, машинный перевод, словарь, оптимальный транспорт, машинное обучение, многоязычный перевод.

Введение

Из-за дискретности текста построение словарного запаса является обязательным условием для нейронного машинного перевода (NMT) и многие другие средства обработки естественного языка (NLP) требуют задачи с использованием нейронных сетей [1]. В настоящее время, подходы к подсловам, такие как кодированные байт-пары (BPE) широко используются в обществе [2] и добиться на практике весьма многообещающих результатов [3]. Ключевой идеей этих подходов является выбор наиболее часто встречающиеся подслов (или части слов с более высокой вероятностью) в качестве словарных токенов.

В теории информации эти частотные подходы являются простыми формами сжатия данных для уменьшения энтропии, что делает корпус легко изучаемым и предсказуемым [4]. Однако влияние размера словарного запаса не в достаточной степени учитывались, поскольку современные подходы рассматривали только частоту (или энтропию) как основные критерии. Многие предыдущие исследования [5] показывают, что словарный запас размер также влияет на производительность нисходящего потока, особенно на малоресурсных задачах. Из-за отсутствия соответствующего индуктивного смещения относительно размера, пробное обучение (а именно обход всех возможных размеров) обычно требует поиска оптимального размера, который требует высоких вычислительных затрат. Для удобства большинство существующих исследований принимают только широко используемые настройки в реализации. Например, 30K-40K является самым популярным параметром размера во всех 42 работах на конференции машинного перевода (WMT) в 2017 и 2018 годах [6].

В этой статье мы предлагаем изучить автоматические вокабулирование путем одновременного рассмотрения энтропии и размера словарного запаса без дорогостоящего пробного обучения. Разработка такого похода вокабулизации необычен по двум основным причинам. Во-первых, сложно найти подходящие цели и одновременно оптимизировать их. Грубо говоря, энтропия корпуса уменьшается с увеличением словарного запаса, что приносит пользу модели обучения [7]. С другой стороны, слишком большое количество токенов приводит к тому,

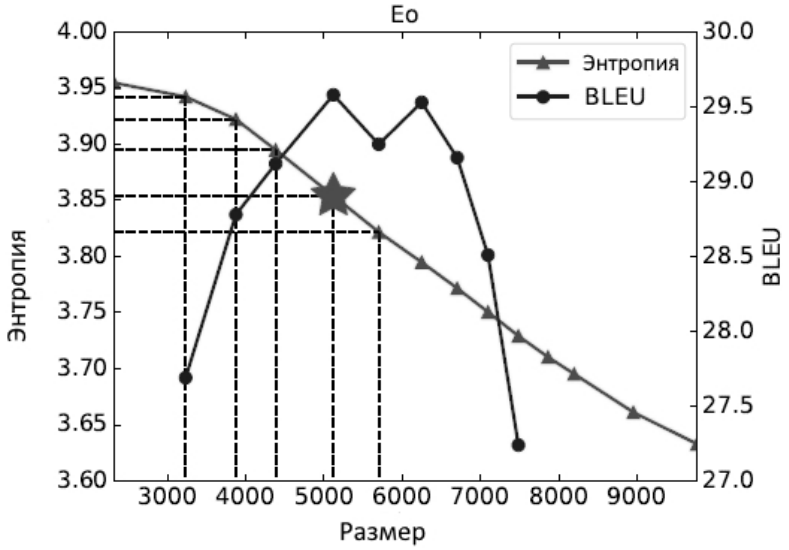

Рис. 1. Иллюстрация предельной полезности. Мы пробуем сгенерированные BPE словари разных размеров для перевода и рисуем их энтропию (см. уравнение 2) и линии BLEU. «Звезда» представляет словарь с максимальной предельной полезностью. Предельная полезность (см. Уравнение 1) оценивает увеличение выгоды (уменьшение энтропии) от увеличения стоимости (размера)

что токены разрежены, что вредит обучению модели [8]. Во-вторых, если предположить, что дано соответствующее измерение, по-прежнему сложно решить такую задачу дискретной оптимизации за счет экспоненциального пространства поиска.

Для решения вышеуказанных проблем мы предлагаем подход к обучению, основанный на изучении словарного запаса через оптимальный Транспорт, сокращенно VNMT. Он может дать соответствующий словарный запас за полиномиальное время, учитывая энтропию корпуса и размер словарного запаса. Конкретно, учитывая вышеупомянутое понимание противоречия между энтропией и размером, мы сначала заимствуем понятие предельной полезности в экономике [5] и предлагают использовать предельную полезность словарного запаса (MUV) в качестве измерения. Понимание довольно простое: в экономике предельная полезность используется для уравновешивания выгод и затрат, а мы используем MUV для балансировки энтропии (преимущества) и размера словарного запаса (стоимости). Высокое MUV ожидается для оптимальности по Парето. Формально MUV определяется как отрицательная производная энтропии к размеру словарного запаса. На рис. 1 приведен пример предельной полезности. Предварительные результаты проверки показывают, что MUV коррелирует с производительностью нисходящего потока в двух третях задач (см. рис. 2).

Тогда наша цель превращается в максимизацию MUV в приемлемой временной сложности. Мы переформулируем нашу дискретную цель оптимизации в проблему оптимального транспорта [7], которую можно решить за полиномиальное время посредством линейного программирования. Интуитивно, процесс вокабулизации можно рассматривать как нахождение оптимальной транспортной матрицы от распределения символов до распределения токенов словарного запаса. Наконец, предлагаемый нами VNMT даст словарь оптимального транспорта матрицы.

Мы оцениваем наш подход по нескольким задачам машинного перевода, включая WMT-14 англо-немецкий перевод, двуязычный перевод TED, и многоязычный перевод TED. Эмпирические результаты показывают, что VNMT превосходит широко используемые словари в разнообразных сценариях. Кроме того, ВОЛЬТ — это легкое решение и не требует дорогостоящих вычислительных ресурсов. Для англо-немецкого перевода, VNMT требуется всего 30 часов использования графического процессора, чтобы найти словари, в то время как традиционное решение BPE-Search занимает 384 часа GPU.

Обзор связанных научных работ

Первоначально большинство нейронных моделей было построено на словарях, состоящих из слов [1]. Эти модели показали многообещающие результаты, но существует общее ограничение, заключающееся в том, что словарям на уровне слов не удается справиться с редкими словами при ограниченном размере словарного запаса.

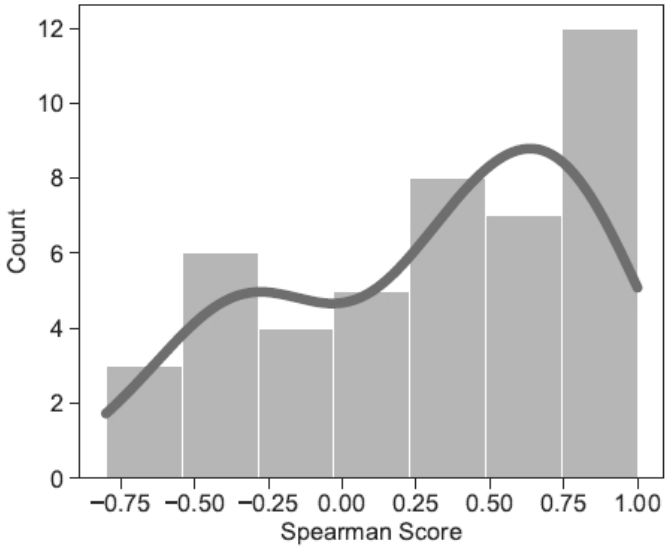

Рис. 2. Производительность MUV и нисходящего потока положительно коррелирует на две трети задач. Ось X классифицирует счёт Спирмена для разных групп. Ось Y показывает количество заданий в каждой группе. Середина — оценка Спирмена — 0,4.

Исследователи недавно предложили несколько передовых подходов к озвучиванию, такие как подход на уровне байтов [4], подход на уровне букв[5], и подсловные подходы [8]. Байт-парное кодирование (BPE) [7] предлагает получить словари уровня подслов. Общая идея такова: объединять пары часто встречающихся последовательностей символов и создавать подсловные единицы. Словари подслов могут рассматриваться как компромисс между словарями на уровне букв и словарями на уровне слов. В сравнении со словарями на уровне слов, это может уменьшить разреженность токенов и увеличить количество общих черт между похожими словами, которые, вероятно, имеют схожие семантические значения, такие как «счастливый» и «счастливее». По сравнению со словарями на уровне букв, у них более короткие предложения без редких слов. Вслед за BPE некоторые варианты недавно были предложены, как BPE-dropout [4], SentencePiece [6] и так далее.

Несмотря на многообещающие результаты, большинство существующих подслов-подходов учитывают только частоту, в то время как влиянием размера словарного запаса пренебрегают. Таким образом, требуется пробная тренировка для подбора оптимального размера, что приводит к высоким вычислительным затратам. Более того, в последнее время некоторые исследования отмечают эту проблему и предлагают некоторые практические решения [10].

Предельная полезность словаря

В этом разделе мы предлагаем найти хорошее измерение словарного запаса с учетом энтропии и размера.

Как показано в разделе 1, нелегко найти соответствующую целевую функцию для их одновременной оптимизации. С одной стороны, с увеличением размера словарного запаса, энтропия корпуса снижается, что приносит пользу обучению модели [8]. С другой стороны, большой словарный запас вызывает взрыв параметров и проблему разреженности токенов, которая мешает обучению модели [9].

Для решения этой проблемы мы заимствуем понятие предельной полезности в экономике [7] и предлагают использовать предельную полезность Вокабуляризации (MUV) как цель оптимизации.

Определение MUV

Формально MUV представляет собой отрицательное отношение энтропии к размеру. Для упрощения мы используем меньший словарный запас для оценки MUV в реализации. Для этого MUV рассчитывается как:

$$\mathcal{M}_{v(k+m)} = \frac{-(\mathcal{H}_{v(k+m)} - \mathcal{H}_{v(k)})}{m}, \quad (1)$$

где $v(k), v(k+m)$ — это два словаря, где k и $k+m$ — это соответствующие токены. $\mathcal{H}_v$ обозначает энтропию корпуса со словарём v, который определяется суммой энтропии токенов. Чтобы избежать влияния длины токена, здесь мы нормализуем энтропию со средней длиной токенов и конечной энтропией, которая определяется как:

$$\mathcal{H}_v = -\frac{1}{l_v}\sum_{j\in v} P(j)\log P(j), \tag{2}$$

где $P(j)$ — соответствующая частотность токена j обучаемого корпуса, а l_v — обычная длина токенов в словаре v.

Предварительные результаты

Для проверки эффективности MUV для измерение словарного запаса, мы проводим эксперименты на 45 языковых парах от TED и рассчитываем показатель корреляции Спирмена* между оценками MUV и BLEU. Две трети пар показывают положительные корреляции, как показано на рисунке 2. Средний балл Спирмена — 0,4. Мы считаем, что это доказывает, что MUV имеет значение.

Максимизация MUV посредством Оптимального Транспорта

В этом разделе подробно описываются предлагаемый подход. Сначала мы опишем оптимальное транспортное решение в разделе 4.1, за которым следует подробности реализации в разделе 4.2.

Вокабулизация с помощью оптимального транспорта

Учитывая набор словарей $\mathbb{V}_{\boldsymbol{S}[t]}$, мы хотим найти словарь с наибольшей энтропией. Как следствие, целевая функция в уравнении 4 становится

$$\min_{v\in\mathbb{V}_{\boldsymbol{S}[t]}} \frac{1}{l_v}\sum_{j\in v} P(j)\log P(j),$$
$$\text{s.t.}\quad P(j) = \frac{\text{Token}(j)}{\sum_{j\in v}\text{Token}(j)},\ l_v = \frac{\sum_{j\in v} len(j)}{|v|}.$$

Токен (j) — это частота токена j в словаре. v. $len(j)$ представляет длину токена j. Обратите внимание, что как распределение P(j), так и средняя длина l_v зависит от выбора v.

Предварительные задачи

Для того, чтобы получить податливую нижнюю границу энтропии, достаточно установить податливую верхняя граница вышеуказанной целевой функции.

Мы принимаем правила слияния для сегментации сырого текста, аналогичного BPE, где два последовательных токена будут объединены в один, если объединенный токен входит в словарный запас. Для этого пусть $\mathbb{T} \in \mathbb{V}_{\boldsymbol{S}[t]}$ будет словарным запасом, содержащим верхние значения S[t] наиболее часто встречающихся токенов, C — множеством символов и $|\mathbb{T}|, |\mathbb{C}|$ будут их размерами соответственно. Поскольку T является элементом $\mathbb{V}_{\boldsymbol{S}[t]}$, ясно, что у нас есть

$$\min_{v\in\mathbb{V}_{\boldsymbol{S}[t]}} \frac{1}{l_v}\sum_{j\in v} P(j)\log P(j) \le \frac{1}{l_{\mathbb{T}}}\sum_{j\in\mathbb{T}} P(j)\log P(j). \tag{5}$$

Здесь мы начинаем с верхней границы приведенной выше целевой функции, то есть

$$\tfrac{1}{l_{\mathbb{T}}}\textstyle\sum_{j\in\mathbb{T}} P(j)\log P(j),$$

а затем найдём уточненный набор токенов от T. Таким образом, мы уменьшаем пространство поиска на подмножества функции T. Пусть P(j; i) — совместное распределение вероятностей токенов и символов, которые мы хотим изучить. Тогда у нас есть

$$\begin{aligned}\sum_{j\in\mathbb{T}} P(j)\log P(j) &= \sum_{j\in\mathbb{T}}\sum_{i\in\mathbb{C}} P(j,i)\log P(j)\\ &= \underbrace{\sum_{j\in\mathbb{T}}\sum_{i\in\mathbb{C}} P(j,i)\log P(j,i)}_{\mathcal{L}_1}\\ &\quad + \underbrace{\sum_{j\in\mathbb{T}}\sum_{i\in\mathbb{C}} P(j,i)(-\log P(i|j))}_{\mathcal{L}_2}.\end{aligned} \tag{6}$$

Детали доказательства можно найти в Приложении C. Поскольку $\mathcal{L}_1$ есть не что иное, как отрицательная энтропия совместного распределения вероятностей $P(j,i)$, обозначим это как $H(P)$. Пусть D — матрица $|\mathbb{C}| \times |\mathbb{T}|$, где (j,i) — вводные данные, которые задаются $\log P(i|j)$, и пусть P совместная матрица вероятностей, то мы можем написать

$$\mathcal{L}_2 = \langle \boldsymbol{P}, \boldsymbol{D}\rangle = \sum_j\sum_i \boldsymbol{P}(j,i)\boldsymbol{D}(j,i). \tag{7}$$

Таким образом, уравнение 6 можно переформулировать как следующую целевую функцию, которая имеет ту же самую форму как целевая функция в оптимальном транспорте:

$$\min_{\boldsymbol{P}\in\mathbb{R}^{m\times n}} \langle \boldsymbol{P}, \boldsymbol{D}\rangle - \gamma H(\boldsymbol{P}). \tag{8}$$

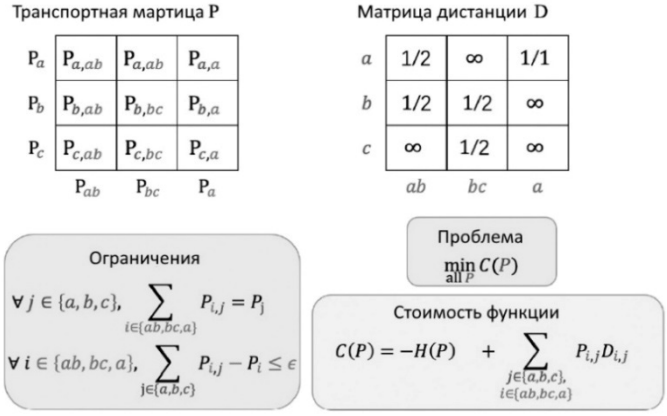

Рис. 4. Детали оптимального транспорта. Цель функции представляет собой сумму отрицательной энтропии и переноса стоимости. Каждый элемент D(i; j) в матрице расстояний является отрицательным логарифмом 1/n, где n — длина токена i. Он определяет расстояние между символом j и токеном i. Чтобы избежать недопустимого транспорта между символом j и токеном i, мы установим расстояние на бесконечность, если целевой токен i не содержат символ j.

Настройка OT

С точки зрения оптимального транспорта, P можно рассматривать как транспортную матрицу, и D можно рассматривать как матрицу расстояний. Интуитивно, оптимальный транспорт заключается в нахождении наилучшей транспортирующей массы из распределения символов к целевому распределению токенов с минимальной работой, это определяется как $\langle \boldsymbol{P}, \boldsymbol{D} \rangle$.

Для проверки обоснованности транспортных решений, добавляем следующие ограничения. Во-первых, чтобы избежать недействительного транспорт между символом i и токеном j, мы установим расстояние на $+\infty$, если целевой токен j не содержит символ i. В противном случае мы используем

$$\frac{1}{len(j)}$$

для оценки $P(i|j)$, где $len(j)$ — длина токена j. Формально матрица расстояний определяется как

$$\boldsymbol{D}(j,i) = \begin{cases} -\log P(i|j) = +\infty, & \text{if } i \notin j \\ -\log P(i|j) = -\log \frac{1}{len(j)}, & \text{otherwise} \end{cases}$$

Кроме того, количество символов фиксировано, и мы устанавливаем сумму каждой строки в транспортной матрице равной вероятность символа i. Верхняя граница требований к символам для каждого токена фиксированы, и мы устанавливаем сумму каждого столбца в транспортной матрице на уровне вероятности токена j. В этом случае ограничения определяются как:

$$|\sum_i \boldsymbol{P}(j,i) - P(j)| \leq \epsilon, \tag{9}$$

и

$$\sum_j \boldsymbol{P}(j,i) = P(i). \tag{10}$$

Учитывая транспортную матрицу P и матрицу расстояний D, конечная цель может быть сформулирована как:

$$\underset{\boldsymbol{P} \in \mathbb{R}^{|\mathbb{C}| \times |\mathbb{T}|}}{\arg\min} \; -H(\boldsymbol{P}) + \langle \boldsymbol{P}, \boldsymbol{D} \rangle,$$

$$\text{s.t.} \quad \sum_i \boldsymbol{P}(i,j) = P(j), \quad |\sum_j \boldsymbol{P}(i,j) - P(i)| \leq \epsilon,$$

с малым $\epsilon > 0$. На рис. 4 показаны детали оптимального транспортного решения. Строго говоря, это несбалансированная энтропийная регуляризованная оптимальная транспортная проблема. Тем не менее, мы все еще можем использовать обобщенный алгоритм Синкхорна, чтобы эффективно найти целевой словарь, как подробно описано в разделе

[7]. Детали алгоритма показаны в алгоритме 1. На каждом временном шаге t мы можем создать новый словарь, связанный с показателями энтропии на основе транспортной матрицы P. Наконец, мы собираем эти словари, связанные с показателями энтропии, и выводим словарный запас, удовлетворяющий уравнению 3.

Применение

Алгоритм 1 описывает процесс VNMT. Первое, мы ранжируем всех кандидатов в токены в соответствии с их частотой). Для упрощения мы принимаем токены, выработанное BPE, (например, BPE-100K) в качестве токенов-кандидатов.

Algorithm 1: VNMT

Input: A sequence of token candidates $\mathbb{L}$ ranked by frequencies, an incremental integer sequence $\boldsymbol{S}$ where the last item of $\boldsymbol{S}$ is less than $|\mathbb{L}|$, a character sequence $\mathbb{C}$, a training corpus D_c

Parameters: u $\in \mathbb{R}_+^{|\mathbb{C}|}$, v $\in \mathbb{R}_+^{|\mathbb{T}|}$

vocabularies = []

for *item in* $\boldsymbol{S}$ **do**

    // Begin of Sinkhorn algorithm

    Initialize u = ones() and v = ones()

    $\mathbb{T} = \mathbb{L}[: item]$

    Calculate token frequencies $P(\mathbb{T})$ based on D_c

    Calculate char frequencies $P(\mathbb{C})$ based on D_c

    Calculate $\boldsymbol{D}$

    while *not converge* **do**

        u = $P(\mathbb{T})/\boldsymbol{D}$v

        v = $P(\mathbb{C})/\boldsymbol{D}^T$u

    optimal_matrix = u.reshape(-1, 1) * $\boldsymbol{D}$ * v.reshape(1, -1)

    // End of Sinkhorn algorithm

    entropy, vocab = get_vocab(optimal_matrix)

    vocabularies.append(entropy,vocab)

Output $\boldsymbol{v}^*$ from vocabularies satisfying Eq. 3

Рис. 5

Мы в этой работе просто принимаем BPE-100K для двуязычного перевода и BPE-300K для многоязычного перевода. Все кандидаты в токены с их вероятностями затем используются для инициализации L в алгоритме 1.

Размер возрастающей целочисленной последовательности S является гиперпараметром и устанавливается в (1K; …; 10K) для двуязычный перевод, (40K; …; 160K) для многоязычного перевода. На каждом временном шаге мы можем получить словарь с максимальной энтропией на основе транспортной матрицы. Неизбежно возникает проблема незаконной перевозки из-за смягчения ограничений.

Мы удаляем токены с распределенными символами, частотность токенов которых менее 0,001. Наконец, мы перечисляем все временные интервалы и выбираем словарный запас, удовлетворяющий уравнению 3 в качестве окончательного словаря.

После создания словаря, VNMT использует жадную стратегию кодирования текста, аналогичную BPE. Для того чтобы кодировать текст, он сначала разбивает предложения на уровень символов-токенов. Затем мы объединяем два последовательных токена в один токен, если объединенный токен находится в словаре. Этот процесс продолжается до тех пор, пока не останется токенов, которые можно объединять. Токены вне словаря будут разделены на более мелкие токены.

Заключение

В этой работе мы предлагаем новый подход к поиску словаря без трейловой подготовки. Вся структура начинается с информационно-теоретического понимания. В соответствии с этим пониманием мы определяем целью вокабуляризации двухэтапную дискретную оптимизацию и предлагаем принципиально оптимальное транспортное решение VNMT.

ЛИТЕРАТУРА

1. Christian Bentz and Dimitrios Alikaniotis. 2016. The word entropy of natural languages. arXiv preprint arXiv:1606.06996.
2. Philip Gage. 1994. A new algorithm for data compression. C Users Journal, 12(2):23–38.
3. ulia Kreutzer and Artem Sokolov. 2018. Learning to segment inputs for NMT favors character-level processing. CoRR, abs/1810.01480.
4. Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020. Very deep transformers for neural machine translation. CoRR, abs/2008.07772.
5. Nathaniel FG Martin and James W England. 2011. Mathematical theory of entropy. 12. Cambridge university press.
6. Nathaniel FG Martin and James W England. 2011. Mathematical theory of entropy. 12. Cambridge university press.
7. Paul A Samuelson. 1937. A note on measurement of utility. The review of economic studies, 4(2):155–161.

Журнал «Современная наука: актуальные проблемы теории и практики»

МИЭТ