УНИВЕРСИТЕТ ИТМО

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2021 Том 21 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2021 Vol. 21 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК УНИВЕРСИТЕТ ИТМО
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2021-21-1-102-108
УДК 004.623

Квантовая семантическая модель поиска текста на арабском языке

Алаа Шакер[1], Игорь Александрович Бессмертный[2], Люсьена Александровна Мирославская[3], Юлия Александровна Королёва[4]✉

[1,2,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] alaashaker11071991@gmail.com, http://orcid.org/0000-0003-2709-0766
[2] bessmertny@itmo.ru, http://orcid.org/0000-0001-6711-6399
[3] lusiena2508@mail.ru, http://orcid.org/0000-0002-6124-7862
[4] jakoroleva@itmo.ru✉, http://orcid.org/0000-0003-1462-1599

Аннотация

Предмет исследования. Рассмотрен процесс извлечения семантики из текстов на арабском языке. Изучена применимость к парам слов теста Белла как мера семантической связанности слов в контексте. Приведены результаты исследования применимости квантового формализма к информационному поиску в текстах на арабском языке. Показано влияние ширины контекста на результативность информационного поиска. **Метод.** Предлагаемая модель поиска базируется на векторном представлении контекста с помощью известного подхода на основе матрицы Hyperspace Analogue to Language (HAL) и теста Белла. Матрица HAL позволяет учитывать частоты встречаемости слов контекста и дистанцию до целевого слова. Использование квантовой теории, оперирующей матрицами плотности вероятностей, позволяет более естественным образом описывать вероятности в векторном представлении слов. **Основные результаты.** Полученные результаты демонстрируют, что использование теста Белла для текстов на арабском языке обеспечивает лучшее ранжирование результатов поисковой выдачи по сравнению с результатами поисковых сервисов. **Практическая значимость.** Результаты исследования могут использоваться при разработке информационно-поисковых систем, а также для дальнейшего развития методов на основе дистрибутивной гипотезы.

Ключевые слова

неравенство Белла, квантовая запутанность, информационный поиск, матрица HAL, алгоритмы информационного поиска, квантовая теория, арабский язык, обработка естественных языков

Ссылка для цитирования: Шакер Алаа, Бессмертный И.А., Мирославская Л.А., Королёва Ю.А. Квантовая семантическая модель поиска текста на арабском языке // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21, № 1. С. 102–108. doi: 10.17586/2226-1494-2021-21-1-102-108

A quantum-like semantic model for text retrieval in Arabic

Alaa Shaker[1], Igor A. Bessmertny[2], Lusiena A. Miroslavskaya[3], Julia A. Koroleva[4]✉

[1,2,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] alaashaker11071991@gmail.com, http://orcid.org/0000-0003-2709-0766
[2] bessmertny@itmo.ru, http://orcid.org/0000-0001-6711-6399
[3] lusiena2508@mail.ru, http://orcid.org/0000-0002-6124-7862
[4] jakoroleva@itmo.ru✉, http://orcid.org/0000-0003-1462-1599

Abstract

The subject of study. The paper focuses on the extraction of semantics from texts in Arabic. In particular, the applicability of the Bell test to word pairs is investigated as a measure of the semantic words relatedness in a context. The study applies the quantum formalism to the task of information retrieval in Arabic texts and presents the results of this work. The authors also examine the influence of the context width on the effectiveness of information retrieval. **Method.** The research is based on the vector representation of the context. It uses the well-known approach based on the HAL (Hyperspace Analogue to Language) matrix and Bell test. The HAL matrix allows taking into account both the frequency of the words occurrence in the context and the distance to the target word. Quantum theory operates with probability density matrices. Quantum theory allows describing probabilities in the vector space in a more natural way,

i.e., words can be represented as vectors. **Main results.** The results demonstrate that using the Bell's test for texts in Arabic provides a better ranking of search results compared to the results of search services. **Practical significance.** The research results can be used in the development of the information retrieval systems, as well as for the further development of methods based on the distributive hypothesis.

Keywords
Bell inequality, quantum entanglement, information retrieval, HAL, IR algorithms, quantum theory, Arabic language, natural language processing

For citation: Shaker A., Bessmertny I.A., Miroslavskaya L.A., Koroleva Ju.A. A quantum-like semantic model for text retrieval in Arabic. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2021, vol. 21, no. 1, pp. 102–108 (in Russian). doi: 10.17586/2226-1494-2021-21-1-102-108

Введение

Прогресс в области технологий информационного поиска, наблюдаемый в течение последних двух десятилетий, обусловлен двумя факторами: заменой синтаксического анализа текстов на статистический и переходом от поиска паттернов в текстах к векторному представлению слов.

Статистический анализ текста базируется на частотах встречаемости слов, позволяет выявлять термины предметной области [1, 2] и строить тезаурусы [3]. Ставший классическим метод TF-IDF (Term Frequency — Inverse Document Frequency) [4] имеет множество модификаций и дополнений, в частности метод взаимной информации [5], использующий данные о совместной встречаемости слов. Наличие множества методов статистического анализа текстов свидетельствует о недостаточной разработанности темы.

Векторное представление слов базируется на дистрибутивной гипотезе Фирта [6] и Харриса [7], согласно которой окружение слова (контекст) несет в себе информацию о семантике слова. Слово здесь представляет собой вектор в пространстве, имеющем размерность словаря. Статистический анализ текстов на основе векторного представления слов обычно осуществляется методами классической (колмогоровской) теории вероятностей. Классическая теория вероятностей базируется на теории множеств и евклидовом многомерном пространстве, в то время как для векторного представления слов более естественным является гильбертово пространство в полярных координатах.

В этой связи заслуживает внимания применение для статистического анализа текстов и векторного представления слов квантовой теории вероятностей, в которой вероятности, как и сами слова, представляются векторами [8]. Авторы провели исследования в области применения квантового формализма для текстов на русском и английском [9], а также китайском языках [10]. В данной работе проведенные ранее исследования распространяются на арабский язык.

Арабский язык имеет некоторые отличия от европейских языков, кроме общеизвестного письма справа налево. В частности, как и в русском языке, глаголы имеют разные формы для мужского и женского рода, а прилагательные склоняются вместе с существительными. Смысл слова находится в зависимости от синтаксической конструкции. Например, слово в разных предложениях может означать «любовь» или «семена». Прилагательные в арабском языке располагаются не перед существительным, а после него. В отличие от других языков, в арабском существительные имеют не только единственное и множественное число, но также двойственное (dual) число. Цель настоящего исследования — оценка универсальности квантового формализма в задачах анализа текстов путем расширения домена за счет арабского языка.

Создание текстового пространства с помощью матрицы HAL

При построении вектора слова в контексте необходимо учитывать как частоту встречаемости окружающих слов, так и расстояние до этого слова. Матрица Hypertext Analog to Language (HAL) позволяет решить данную проблему за счет того, что для каждого i-го слова в элементах матрицы накапливается величина $S–d_{ij} + 1$, где d_{ij} — расстояние от i-го до j-го слова в окне размером S [11, 12]. Таким образом, формируется квадратная матрица, имеющая размерность словаря по каждой координате [13]. Также HAL-матрица является чувствительной к порядку слов, что помогает получить правильное представление о контексте текста, например, в двух утверждениях «Маркс критиковал экономистов» и «Экономисты критиковали Маркса». Запрос «Маркс» «критиковал» и «экономисты» коммутативен, тогда как семантика не кажется коммутативной [14, 15].

На значения векторов HAL влияет размер окна: более широкое окно означает большую вероятность ассоциации между двумя терминами, но большой размер может быть непоказательным при недостаточном соответствии. С другой стороны, маленький размер окна означает сильную связь между двумя терминами, но также может быть неустойчивым показателем при многократном переобучении [13].

Метод векторизации текста. Квантовая теория вероятностей является геометрическим расширением колмогоровской теории вероятностей, поэтому опишем семантическое пространство документа в геометрических терминах и определим базисный вектор, по которому будут генерироваться остальные векторы. В N-мерном пространстве каждый документ будет иметь связанный вектор. Состояние вектора документа — это сумма всех содержащихся в нем векторов слов $|\mathbf{W}_i\rangle$, которые он содержит. Состояние каждого вектора слова может быть извлечено из строк симметричной матрицы HAL. Определим состояние вектора документа следующим образом:

$$|\psi\rangle = \sum_{i}^{N} |\mathbf{W}_i\rangle.$$

Выясним, как два слова связаны в документе. Для этого возьмем две строки из матрицы HAL, относящиеся к словам A и B. Представим эти два слова в виде $\{|\mathbf{W}_A\rangle, |\mathbf{W}_B\rangle\}$, которые будут рассмотрены как базисные векторы.

Применим процесс ортогонализации Грама–Шмидта к неортогональному базису $\{|\mathbf{W}_A\rangle, |\mathbf{W}_B\rangle\}$ и $\{|\mathbf{W}_B\rangle, |\mathbf{W}_A\rangle\}$, в результате получим две координаты для первого слова $|\mathbf{u}_A\rangle$, $|\mathbf{u}_{A_\perp}\rangle$ и для второго $|\mathbf{u}_B\rangle$, $|\mathbf{u}_{B_\perp}\rangle$. Символ «$\perp$» означает, что угол между вектором $|\mathbf{u}_A\rangle$ и вектором $|\mathbf{u}_{A_\perp}\rangle$ равен 90°, другими словами, рассматриваемые вектора ортогональны.

Теперь можно выполнить операцию проекции вектора документа $|\psi\rangle$ на данную ортогональную основу. Для этого запишем вектор всего документа в виде:

$$|\psi\rangle = a|\mathbf{u}_A\rangle + b|\mathbf{u}_{A_\perp}\rangle = c|\mathbf{u}_B\rangle + d|\mathbf{u}_{B_\perp}\rangle,$$

где b, c и d — проекции вектора документа на базис рассматриваемого слова «a». Другими словами, b — проекция вектора документа $|\psi\rangle$ на базис $|\mathbf{u}_{A_\perp}\rangle$, c — проекция вектора документа $|\psi\rangle$ на базис $|\mathbf{u}_B\rangle$, d — проекция вектора документа $|\psi\rangle$ на базис $|\mathbf{u}_{B_\perp}\rangle$.

Коэффициенты базисных векторов (b, c и d) могут быть вычислены путем проецирования вектора документа на базисный вектор, например, коэффициент a вектора $|\mathbf{u}_A\rangle$ может быть вычислен:

$$a = \frac{\langle \mathbf{u}_A|\psi\rangle}{\sqrt{\langle \mathbf{u}_A|\psi\rangle^2 + \langle \mathbf{u}_{A_\perp}|\psi\rangle^2}}.$$

Тест Белла. Тест на основе неравенства Белла используется в физике для определения наличия запутанности между двумя квантовыми частицами. В данной работе с помощью теста Белла проанализирована связь между двумя словами в тексте. Семантическое пространство сформировано с использованием матрицы HAL [14].

Тест Белла в абстрактной форме представлен формулой:

$$S_{bell} = |E(A, B) - E(A, C)| + |E(B, D) + E(C, D)|, \quad (1)$$

где A, B, C и D — исходы теста; $E(X, Y)$ — коэффициенты корреляции результата взаимных тестов X и Y.

Экспериментально с фотонами может быть получен случай $2 \le S_{bell} \le 2\sqrt{2}$. Данный результат достигается с квантовыми запутанными состояниями. В меньшей степени рассмотрен случай, когда $2 \le S_{bell} \le 2\sqrt{2} S_{bell} > 2\sqrt{2}$. Данный случай также известен как граница Цирельсона [16, 17]. Зона между $2\sqrt{2}$ и 4 называется областью «без сигнала». Максимальное значение $S_{bell} = 4$ получено с помощью логических вероятностных конструкций, часто называемых блоками PR (Popescu and Rohrlich) [16]. Область меньше 2 означает, что между двумя частицами нет состояния сцепления.

Операторы запросов. Назначение операторов запросов — количественная оценка запроса в рамках используемого формализма. Операторы запроса возвращают +1, если содержание документа соответствует запросу, и –1 в ортогональном направлении. Операторы будут использовать спин-матрицу Паули, которая выглядит следующим образом:

$$\hat{\mathbf{A}} = \hat{\mathbf{B}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Данные операторы ассоциированы с наблюдаемыми следующим образом:

$$\hat{\mathbf{A}}|\psi\rangle = a|\mathbf{u}_A\rangle - b|\mathbf{u}_{A_\perp}\rangle, \hat{\mathbf{B}}|\psi\rangle = c|\mathbf{u}_B\rangle - d|\mathbf{u}_{B_\perp}\rangle. \quad (2)$$

Значения ожиданий операторов из формулы (2) рассчитываются так же, как и в квантовой механике, с помощью правила Борна. Например, среднее значение в контексте документа, связанного с i-м документом ($|\psi i\rangle$) для запроса об A, можно записать в виде:

$$\langle \mathbf{A} \rangle_\psi = \langle \psi|\hat{\mathbf{A}}|\psi\rangle = 2a^2 - 1.$$

Значения оценок варьируются от +1 до –1. Значение +1 может быть получено, когда вектор документа коллинеарен вектору запроса, и –1, когда он ортогонален.

Другие операторы могут быть определены с помощью, например, матрицы Паули, которая имеет вид:

$$\hat{\mathbf{A}}_x = \hat{\mathbf{B}}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (3)$$

Применив оператор (3) к вектору документа, получим:

$$\hat{\mathbf{A}}_x|\psi\rangle = b|\mathbf{u}_A\rangle + a|\mathbf{u}_{A_\perp}\rangle, \hat{\mathbf{B}}_x|\psi\rangle = d|\mathbf{u}_B\rangle + c|\mathbf{u}_{B_\perp}\rangle.$$

Матрица $\hat{\mathbf{A}}$ представляет собой матрицу вращения оператора Паули. Принято использовать три матрицы вращения Паули $\hat{\mathbf{A}}$, $\hat{\mathbf{A}}_x$ и $\hat{\mathbf{A}}_y$ для определения осей вращения.

Данный оператор переключает компоненты состояния вектора. Результат можно интерпретировать как меру различного значения в документе по отношению к исходному направлению, соответствующему слову A [18].

Матрицы HAL всегда содержат вещественные числа, поэтому расчет на основе комплексных чисел не требуется, и спиновая матрица Паули

$$\hat{\mathbf{A}}_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

не используется.

Объединение операторов и значений ожиданий

Выполнять все вычисления наиболее удобно на основе одного базиса, а именно слова A, показанного в уравнении $|\mathbf{u}_A\rangle$, $|\mathbf{u}_{A_\perp}\rangle$. Для этого преобразуем операцию ($\hat{\mathbf{B}}$, $\hat{\mathbf{B}}_x$) из базиса $|\mathbf{u}_B\rangle$, $|\mathbf{u}_{B_\perp}\rangle$ в базис слова A $|\mathbf{u}_A\rangle$, $|\mathbf{u}_{A_\perp}\rangle$ и матрицу $\mathbf{M}$ из $|\mathbf{u}_B\rangle$, $|\mathbf{u}_{B_\perp}\rangle$ в $|\mathbf{u}_A\rangle$, $|\mathbf{u}_{A_\perp}\rangle$. Введем новое обозначение $p = \langle \mathbf{u}_B|\mathbf{u}_A\rangle = \langle \mathbf{u}_{B_\perp}|\mathbf{u}_{A_\perp}\rangle$ и запишем матрицу $\mathbf{M}$ в упрощенном виде:

$$\mathbf{M} = \begin{pmatrix} \langle \mathbf{u}_B|\mathbf{u}_A\rangle & \langle \mathbf{u}_B|\mathbf{u}_{A_\perp}\rangle \\ \langle \mathbf{u}_{B_\perp}|\mathbf{u}_A\rangle & \langle \mathbf{u}_{B_\perp}|\mathbf{u}_{A_\perp}\rangle \end{pmatrix} = \begin{pmatrix} p & \sqrt{1-p^2} \\ -\sqrt{1-p^2} & p \end{pmatrix}. \quad (4)$$

Таким образом, любой оператор, выраженный в его матричной форме, в базисе, ассоциированном со словом B, может быть записан в базисе, ассоциированном со словом A, с помощью матрицы преобразования $\mathbf{M}$.

Из определения (4) матричная форма операции ($\hat{\mathbf{B}}$, $\hat{\mathbf{B}}_x$) в базисе, связанном со словом A, можно записать в виде:

$$\hat{\mathbf{B}} = \mathbf{M^{-1}}\cdot\hat{\mathbf{A}}\cdot\mathbf{M} = \begin{pmatrix} 2p^2-1 & 2p\sqrt{1-p^2} \\ 2\sqrt{1-p^2} & 1-2p^2 \end{pmatrix},$$

$$\hat{\mathbf{B}}_x = \mathbf{M^{-1}}\cdot\hat{\mathbf{A}}_x\cdot\mathbf{M} = \begin{pmatrix} -2p\sqrt{1-p^2} & 2p^2-1 \\ 2p^2-1 & 2p\sqrt{1-p^2} \end{pmatrix}.$$

Расчет теста Белла. Для определения степени, в которой документ соответствует слову A и слову B одновременно $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, можно использовать формулу $\langle\hat{\mathbf{A}}\hat{\mathbf{B}}\rangle_\psi = \langle\psi|\hat{\mathbf{A}}\hat{\mathbf{B}}|\psi\rangle$.

Вычислим квантовое среднее, определяемое в формуле (1). Для этого используем различные операторы запросов, которые могут рассматриваться как измерительные устройства, а затем определим параметр запроса Белла:

$$S_{query} = |\langle\hat{\mathbf{A}}\hat{\mathbf{B}}_+\rangle_\psi + \langle\hat{\mathbf{A}}_x\hat{\mathbf{B}}_+\rangle_\psi| + |\langle\hat{\mathbf{A}}\hat{\mathbf{B}}_-\rangle_\psi - \langle\hat{\mathbf{A}}_x\hat{\mathbf{B}}_-\rangle_\psi|,$$

где операторы представлены в следующем виде:

$$\hat{\mathbf{A}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},\ \hat{\mathbf{A}}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \hat{\mathbf{B}}_+ = -\frac{B+B_x}{\sqrt{2}},\ \hat{\mathbf{B}}_- = \frac{B-B_x}{\sqrt{2}}.$$

Результат экспериментов

В качестве исходных данных использованы результаты выполнения поисковых запросов в Google. Поисковый запрос состоял из двух слов на арабском языке. Из поисковой выдачи были отобраны тексты, написанные на классическом арабском языке. Эксперименты были проведены для двух тем: «информационная инженерия» и «сельскохозяйственная инженерия». В экспериментальной части рассмотрены три ситуации:

1) если в тексте нет ни первого слова запроса, ни второго;
2) если есть одно слово;
3) если оба слова встречаются в текстах.

Первый запрос на арабском языке выглядел так: «الهندسة الزراعية»[1], что означает на русском языке «сельскохозяйственная инженерия». Результаты применения теста Белла с разным размером окна матрицы HAL представлены на рис. 1.

В тексте, озаглавленном «исторические исследования» (на арабском языке دراسا تاريخية)[2], не содержится ни одного слова запроса, поэтому результат теста Белла A будет равен нулю.

Второй текст, помеченный как «инженерия» (на арабском языке الهندسة)[3], имеет самый высокий показатель запутанности. Данные результаты вызывают наибольшее количество вопросов и должны быть рассмотрены с логической и математической стороны. С логической стороны вектор документа будет иметь максимальную запутанность с вектором существующего слова при исчезновении второго слова, т. е. первое слово будет иметь связь с неизвестным словом «неоднозначное состояние».

С математической точки зрения, если второе слово не встретилось в тексте, это означает, что вектор второго слова будет нулевым, а вектор первого – не нулевым, то вектор документа может быть представлен следующим способом:

$$a = \frac{\langle\mathbf{u}_A|\psi\rangle}{\sqrt{\langle\mathbf{u}_A|\psi\rangle^2 + \langle\mathbf{u}_{A_\perp}|\psi\rangle^2}}\ \text{Вопрос}\ \frac{\langle\mathbf{u}_A|\psi\rangle}{\sqrt{\langle\mathbf{u}_A|\psi\rangle^2}} = 1,\ \langle\mathbf{u}_{A_\perp}|\psi\rangle = 0,$$

cause $u_{A_\perp}$ *is zero*,

$$b = \frac{\langle\mathbf{u}_{A_\perp}|\psi\rangle}{\sqrt{\langle\mathbf{u}_A|\psi\rangle^2 + \langle\mathbf{u}_{A_\perp}|\psi\rangle^2}}\ \text{Вопрос 0},\ \langle\mathbf{u}_{A_\perp}|\psi\rangle = 0,$$

cause $u_{A_\perp}$ *is zero*

$$\Rightarrow |\psi\rangle = a|\mathbf{u}_A\rangle + b|\mathbf{u}_{A_\perp}\rangle = 1|\mathbf{u}_A\rangle + 0|\mathbf{u}_{A_\perp}\rangle,$$

$$M = \begin{pmatrix} p & \sqrt{1-p^2} \\ -\sqrt{1-p^2} & p \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

$$\hat{\mathbf{B}} = \mathbf{M^{-1}}\cdot\hat{\mathbf{A}}\cdot\mathbf{M} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\hat{\mathbf{B}}_x = \mathbf{M^{-1}}\cdot\hat{\mathbf{A}}_x\cdot\mathbf{M} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix},$$

$$\mathbf{B}_+ = -\frac{\mathbf{B}+\mathbf{B}_x}{\sqrt{2}} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\hat{\mathbf{B}}_- = \frac{\mathbf{B}-\mathbf{B}_x}{\sqrt{2}} = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix},$$

$$\hat{\mathbf{A}}\hat{\mathbf{B}}_+ = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},\ \hat{\mathbf{A}}_x\hat{\mathbf{B}}_+ = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

$$\hat{\mathbf{A}}\hat{\mathbf{B}}_- = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix},\ \hat{\mathbf{A}}_x\hat{\mathbf{B}}_- = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

$$\langle\hat{\mathbf{A}}\hat{\mathbf{B}}_+\rangle_\psi = \langle\psi|\hat{\mathbf{A}}\hat{\mathbf{B}}_+|\psi\rangle = \frac{1}{\sqrt{2}}[1\quad 0]\cdot\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\cdot\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}}.$$

Таким же образом, $\langle\hat{\mathbf{A}}_x\hat{\mathbf{B}}_+\rangle_\psi = \frac{1}{\sqrt{2}}$, $\langle\hat{\mathbf{A}}\hat{\mathbf{B}}_-\rangle_\psi = -\frac{1}{\sqrt{2}}$, $\langle\hat{\mathbf{A}}_x\hat{\mathbf{B}}_-\rangle_\psi = \frac{1}{\sqrt{2}}$,

$$S_{query} = \left|\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right| + \left|-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}\right| = 2\sqrt{2}.$$

Данное состояния не существовало в физике, потому что все эксперименты проводились с двумя частицами, в рассматриваемом случае подразумевается, что одна частица обязательно имеет запутанность.

Остальные три текста, обозначенные как «сельскохозяйственная инженерия», «сельскохозяйственная инженерия 2»[4] и «сельскохозяйственная инженерия 3»[5]

[1] Сельскохозяйственная инженерия [Электронный ресурс]. Режим доступа: https://www.marefa.org/هندسة_زراعية (дата обращения: 14.12.2020).

[2] Исторические исследования [Электронный ресурс]. Режим доступа: https://ar.wikipedia.org/wiki/تاريخ (дата обращения: 14.12.2020).

[3] Инженерия [Электронный ресурс]. Режим доступа: https://mawdoo3.com/ما_هي_الهندسة (дата обращения: 14.12.2020).

[4] Сельскохозяйственная инженерия 2 [Электронный ресурс]. Режим доступа: https://www.easyunime.com/advice/تخصص-الهندسة-الزراعية-2803/ (дата обращения: 14.12.2020).

[5] Сельскохозяйственная инженерия 3 [Электронный ресурс]. Режим доступа: https://ar.wikipedia.org/wiki/هندسة_زراعية (дата обращения: 14.12.2020).

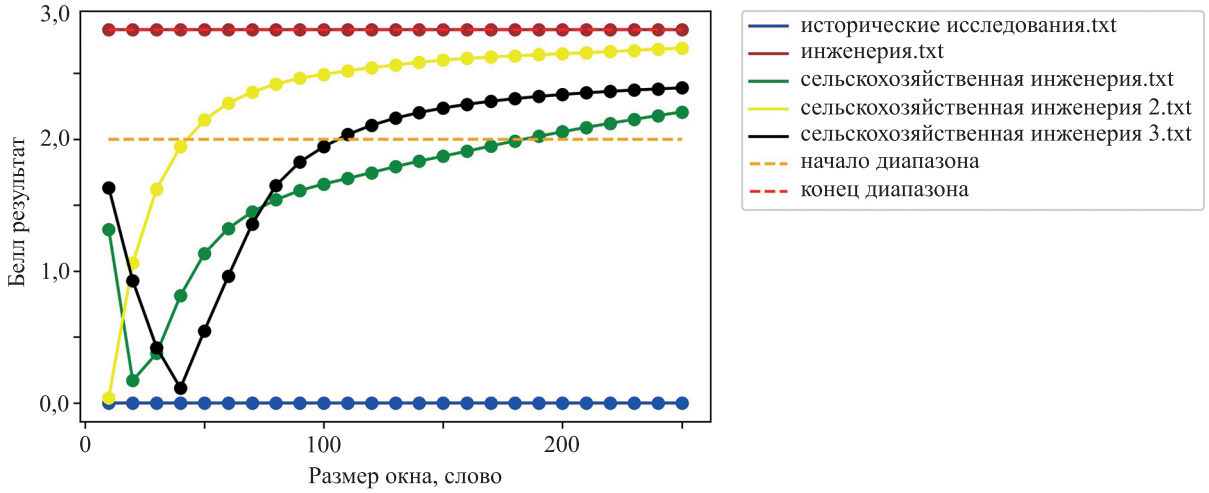

Рис. 1. Результат применения теста Белла на примере запроса «сельскохозяйственная инженерия»

Fig. 1. The result of the query “agricultural engineering” (in Arabic “الهندسة الزراعية”)

содержат оба слова запроса с разными результатами теста Белла.

Текст с названием «сельскохозяйственная инженерия» (на арабском языке الهندسة الزراعية) имеет более низкую оценку, чем два других текста. Это связано с тем, что большая его часть рассказывает об истории сельского хозяйства и о том, как оно развивалось, поэтому текст отображается на пятом месте в результатах поиска. В поисковой системе Google текст с названием «сельскохозяйственная инженерия 3» появляется на первом месте среди результатов поиска, хотя он очень короткий и не содержит достаточно информации по теме.

Третий текст, обозначенный как «сельскохозяйственная инженерия 2» имеет большую взаимосвязь, чем два остальных текста, но он появляется лишь на второй странице результатов поиска в Google. Этот текст содержит достаточно информации, например: информацию о специализации сельскохозяйственного машиностроения и ее важности, о дисциплинах и академических предметах, об университетах, предлагающих сельскохозяйственную инженерию и областях их работы. Таким образом, можно сделать вывод о том, что можно полагаться на оценку теста Белла для оценки степени соответствия текста по интересующему предмету поиска.

Результаты применения теста Белла к запросу на арабском языке «الهندسة المعلوماتية»[1], означающий «информационная инженерия», с разным размером окна матрицы HAL показаны на рис. 2.

В тексте под названием «исторические исследования» (на арабском языке دراسات تاريخية) нет ни одного слова запроса, поэтому результат был нулевым, текст под названием «инженерия» (на арабском языке الهندسة) содержит только одно слово «инженерия» из двух слов запроса, остальные четыре текста содержат оба слова. Результаты получены в диапазоне размеров окна [80–150], чтобы избежать попадания в две основные ситуации: недостаточное соответствие «маленького размера окна» или переобучение «большого размера окна».

В тексте под названием «информационная инженерия 4» с сайта «Wikipedia» (на арабском языке الهندسة الزراعية)[2] говорится о факультете информационной инженерии в Дамасском университете (присвоенные степени, названия предметов, факультеты и специальности). Это означает, что не хватает информации, которая ищется. Текст появляется на втором месте в результатах поисковой выдачи Google, а оценка теста Белла была получена меньше 2.

Текст под названием «информационная инженерия» соответствует теме поиска, например, определение информационной инженерии, истории, академического содержания, областей работы и применения. Полученный текст помещен на первое место в поиске Google, и он имеет оценку теста Белла в пределах $2{,}2\sqrt{2}$, а это говорит о наличии запутанности между двумя словами запроса в тексте.

Другой текст под названием «информационная инженерия 2»[3], имеет средний размер, содержит 1034 слова и является в достаточной степени содержательным. Результат теста Белла лежит в диапазоне [2; 2,5], что указывает на то, что текст имеет отношение к предмету поиска. В то же время этот текст находится на 12-м месте в поиске Google.

Последний текст под названием «информационная инженерия 3»[4] содержит определение информационной инженерии, областей работы и ее специализаций, а также некоторые другие небольшие вопросы с ответами,

[1] Информационная инженерия [Электронный ресурс]. Режим доступа: https://www.marefa.org/الهندسة_المعلوماتية (дата обращения: 14.12.2020).

[2] Информационная инженерия 4 [Электронный ресурс]. Режим доступа: https://ar.wikipedia.org/wiki/كلية_الهندسة_المعلوماتية_بجامعة_دمشق (дата обращения: 14.12.2020).

[3] Информационная инженерия 2 [Электронный ресурс]. Режим доступа: https://khatwa-sy.com/الهندسة-المعلوماتية.html (дата обращения: 14.12.2020).

[4] Информационная инженерия 3 [Электронный ресурс]. Режим доступа: http://damascusuniversity.edu.sy/ite/index.php?lang=1&set=5&id=3 (дата обращения: 14.12.2020).

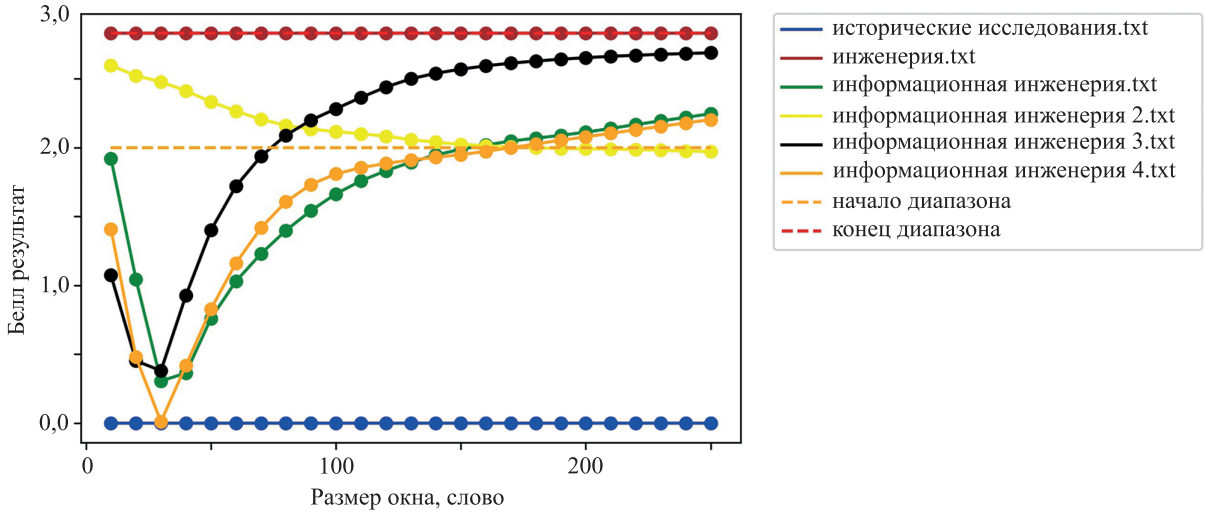

Рис. 2. Результат запроса «информационная инженерия»

Fig. 2. The result of the query “information engineering” (in Arabic “الهندسة المعلوماتية”)

которые были бы интересны пользователю. Результат теста Белла находится в диапазоне [2; 2,7], что говорит о наличии запутанности между двумя рассматриваемыми словами в данном тексте, т. е. семантической связи в контексте документа.

Заключение

Проведенные исследования продемонстрировали, что дистрибутивная гипотеза Харриса и метод Hyperspace Analogue to Language (HAL) дают возможность построить семантическое пространство текста не только для европейских языков, но и для арабского языка. В результате выполненных исследований обнаружено, что параметр Белла при анализе арабских текстов сильно зависит от размера окна HAL, как это имело место и для других языков. На основе полученных результатов можно предположить, что для данного типа модели существует оптимальный размер окна, который максимизирует параметр Белла.

Предложенный подход можно использовать для улучшения поиска релевантных текстов путем ранжирования результатов поиска с использованием теста Белла. Таким образом, можно объединить два алгоритма: сначала использовать традиционный статистический метод (TF-IDF) для получения списка текстов-кандидатов, затем, использовать квантовоподобную семантическую модель. В результате это позволит упорядочить файлы по убыванию релевантности запросу. Для того чтобы избежать попадания в ситуацию, когда одно из слов запроса отсутствует в тексте, необходимо отфильтровывать такие тексты на первом этапе, либо рассматривать только результаты, попадающие в диапазон $2 \le S_{bell} < 2\sqrt{2}$ теста Белла.

Литература

1. Yang Y., Pedersen J.O. A comparative study on feature selection in text categorization // ICML’97: Proc. of the Fourteenth International Conference on Machine Learning. 1997. P. 412–420.
2. Peñas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access // Proc. of the Corpus Linguistics 2001 Conference. 2001. P. 458–465.
3. Бессмертный И.А., Нугуманова А.Б. Метод автоматического построения тезаурусов на основе статистической обработки текстов на естественном языке // Известия Томского политехнического университета. 2012. Т. 321. № 5. С. 125–130.
4. Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. V. 60. N 5. P. 493–502. doi: 10.1108/00220410410560573
5. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information // Information Systems Frontiers. 2011. V. 13. N 1. P. 115–125. doi: 10.1007/s10796-010-9278-5
6. Harris Z.S. Distributional structure // Word. 1954. V. 10. N 2-3. P. 146–162. doi: 10.1080/00437956.1954.11659520
7. Sahlgren M. The distributional hypothesis // Rivista di Linguistica. 2008. V. 20. N 1. P. 33–53.
8. Melucci M., Piwowarski B. Quantum mechanics and information retrieval: From theory to application // Proc. 4th International Conference on the Theory of Information Retrieval, ICTIR 2013.

References

1. Yang Y., Pedersen J.O. A comparative study on feature selection in text categorization. *ICML’97: Proc. of the Fourteenth International Conference on Machine Learning*, 1997, pp. 412–420.
2. Peñas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access. *Proc. of the Corpus Linguistics 2001 Conference*, 2001, pp. 458–465.
3. Bessmertnyi I.A., Nugumanova A.B. Method for automatic construction of thesauri based on statistical processing of natural language texts. *Bulletin of the Tomsk Polytechnic University*, 2012, vol. 321, no. 5, pp. 125–130. (in Russian)
4. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 2004, vol. 60, no. 5, pp. 493–502. doi: 10.1108/00220410410560573
5. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information. *Information Systems Frontiers*, 2011, vol. 13, no. 1, pp. 115–125. doi: 10.1007/s10796-010-9278-5
6. Harris Z.S. Distributional structure. *Word*, 1954, vol. 10, no. 2-3, pp. 146–162. doi: 10.1080/00437956.1954.11659520
7. Sahlgren M. The distributional hypothesis. *Rivista di Linguistica*, 2008, vol. 20, no. 1, pp. 33–53.
8. Melucci M., Piwowarski B. Quantum mechanics and information retrieval: From theory to application. *Proc. 4th International Conference on the Theory of Information Retrieval, ICTIR 2013*,

2013. P. 1. (ACM International Conference Proceeding Series). doi: 10.1145/2499178.2499202
9. Trukhanov A., Platonov A., Bessmertny I. Using quantum probability for word embedding problem // CEUR Workshop Proceedings. 2020. V. 2590.
10. Bessmertny I.A., Huang X., Platonov A.V., Yu C., Koroleva J.A. Applying the Bell's test to chinese texts // Entropy. 2020. V. 22. N 3. P. 275. doi: 10.3390/e22030275
11. Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behavior Research Methods, Instruments, and Computers. 1996. V. 28. N 2. P. 203–208. doi: 10.3758/BF03204766
12. Galofaro F., Toffano Z., Doan B.-L. A quantum-based semiotic model for textual semantics // Kybernetes. 2018. V. 47. N 2. P. 307–320. doi: 10.1108/K-05-2017-0187
13. Шакер А. Using bell test for realizing a quantum-like semantic model for text retrieval in arabic texts // Сборник тезисов докладов конгресса молодых ученых. 2020 [Электронный ресурс]. URL: https://kmu.itmo.ru/digests/article/4084. IET — 2020 (дата обращения: 14.12.2020).
14. Galofaro F., Doan B.-L., Toffano Z. Linguistics and quantum theory: epistemological perspectives // Proc. 19th IEEE International Conference on Computational Science and Engineering, 14th IEEE International Conference on Embedded and Ubiquitous Computing and 15th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2016. P. 660–667. doi: 10.1109/CSE-EUC-DCABES.2016.257
15. Kartsaklis D. Compositional operators in distributional semantics // Springer Science Reviews. 2014. V. 2. N 1-2. P. 161–177. doi: 10.1007/s40362-014-0017-z
16. Cabello A. Violating Bell's inequality beyond Cirel'son's bound // Physical Review Letters. 2002. V. 88. N 6. P. 060403. doi: 10.1103/PhysRevLett.88.060403
17. Popescu S., Rohrlich D. Quantum nonlocality as an axiom // Foundations of Physics. 1994. V. 24. N 3. P. 379–385. doi: 10.1007/BF02058098
18. Bruza P.D., Woods J. Quantum collapse in semantic space: interpreting natural language argumentation // Proc. 2nd Quantum Interaction Symposium. 2008. P. 141–147.

2013, pp. 1. ACM International Conference Proceeding Series. doi: 10.1145/2499178.2499202
9. Trukhanov A., Platonov A., Bessmertny I. Using quantum probability for word embedding problem. *CEUR Workshop Proceedings*, 2020, vol. 2590.
10. Bessmertny I.A., Huang X., Platonov A.V., Yu C., Koroleva J.A. Applying the Bell's test to chinese texts. *Entropy*, 2020, vol. 22, no. 3, pp. 275. doi: 10.3390/e22030275
11. Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 1996, vol. 28, no. 2, pp. 203–208. doi: 10.3758/BF03204766
12. Galofaro F., Toffano Z., Doan B.-L. A quantum-based semiotic model for textual semantics. *Kybernetes*, 2018, vol. 47, no. 2, pp. 307–320. doi: 10.1108/K-05-2017-0187
13. Shaker A. Using bell test for realizing a quantum-like semantic model for text retrieval in arabic texts. *Collection of Abstracts of the Congress of Young Scientists*, 2020. Available at: https://kmu.itmo.ru/digests/article/4084. IET — 2020 (accessed: 14.12.2020).
14. Galofaro F., Doan B.-L., Toffano Z. Linguistics and quantum theory: epistemological perspectives. *Proc. 19th IEEE International Conference on Computational Science and Engineering, 14th IEEE International Conference on Embedded and Ubiquitous Computing and 15th International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 2016, pp. 660–667. doi: 10.1109/CSE-EUC-DCABES.2016.257
15. Kartsaklis D. Compositional operators in distributional semantics. *Springer Science Reviews*, 2014, vol. 2, no. 1-2, pp. 161–177. doi: 10.1007/s40362-014-0017-z
16. Cabello A. Violating Bell's inequality beyond Cirel'son's bound. *Physical Review Letters*, 2002, vol. 88, no. 6, pp. 060403. doi: 10.1103/PhysRevLett.88.060403
17. Popescu S., Rohrlich D. Quantum nonlocality as an axiom. *Foundations of Physics*, 1994, vol. 24, no. 3, pp. 379–385. doi: 10.1007/BF02058098
18. Bruza P.D., Woods J. Quantum collapse in semantic space: interpreting natural language argumentation. *Proc. 2nd Quantum Interaction Symposium*, 2008, pp. 141–147.

Авторы

Шакер Алаа — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, alaashaker11071991@gmail.com, http://orcid.org/0000-0003-2709-0766
Бессмертный Игорь Александрович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 36661767800, bessmertny@itmo.ru, http://orcid.org/0000-0001-6711-6399
Мирославская Люсьена Александровна — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, lusiena2508@mail.ru, http://orcid.org/0000-0002-6124-7862
Королёва Юлия Александровна — кандидат технических наук, преподаватель, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57216157310, jakoroleva@itmo.ru, http://orcid.org/0000-0003-1462-1599

Authors

Alaa Shaker — Postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, alaashaker11071991@gmail.com, http://orcid.org/0000-0003-2709-0766
Igor A. Bessmertny — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 36661767800, bessmertny@itmo.ru, http://orcid.org/0000-0001-6711-6399

Lusiena A. Miroslavskaya — Postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, lusiena2508@mail.ru, http://orcid.org/0000-0002-6124-7862
Julia A. Koroleva — PhD, Lecturer, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57216157310, jakoroleva@itmo.ru, http://orcid.org/0000-0003-1462-1599

Статья поступила в редакцию 16.11.2020
Одобрена после рецензирования 20.12.2020
Принята к печати 05.02.2021

Received 16.11.2020
Approved after reviewing 20.12.2020
Accepted 05.02.2021