The Effectiveness of Using Bell Inequality Test for Information Retrieval in Arabic Texts

Alaa Shaker[a], Alaa Aldarf[a] and Prof Igor Alexandrovich Bessmertny[a]

[a]ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

Abstract

In information retrieval, getting the most relevant documents to the target topic on the top places of the search results and evaluating a query's effectiveness are essential tasks. This article suggests a new method for searching texts in the Arabic language by using a quantum-like semantic model. Also, it shows the analogy between natural language texts and quantum-like systems by using Bell's test. The HAL (Hypertext analog to language) matrix is applied for building the quantum-space. The effect of the window size of the HAL matrix is also a subject of research. In this work, three experiments are discussed. The first one is applying semantic-quantum to many Arabic texts and the applicability to use this model in natural language processing. In the second experiment, the model will be checked for its ability to find the correlation between titles and texts through a dataset consisting of five-hundred Arabic texts. Our model will be applied in the third experiment to retrieve data from the Arabic dataset consisting of a hundred texts with ten queries and evaluate the retrieved data by three criteria (precision, recall, and F-measure). This approach provides a promising way for information retrieval in Arabic texts, depending on the relation between words of texts. That will help to order retrieved files to get the relevant files to a query in the top places, and that will increase the efficiency of the search engine.

Keywords

Bell inequality, entanglement, information retrieval, Hyperspace analog language (HAL), Quantum Theory, Arabic language, Natural language processing

1. Introduction

Information retrieval (IR) is obtaining information resources relevant to an information need from a collection of those resources. The increasing number of internet resources contradicts with opportunities for effective search. Traditional search engines use statistical algorithms "like TF-IDF". These algorithms base on the statistics of words in documents that cannot always be efficient because they do not take into account the meaning of the query and texts. In contrary, humans do not care about the number of words in the document; the most essential is finding what we search for, more precisely, the document's meaning. Here is a serious problem and the goal of this work is to find how computers can get the meaning or sense of texts. Although,

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10–11, 2020, Online Saint Petersburg, Russia

eng.alaashaker4@gmail.com (A. Shaker); alaadarf2020@gmail.com (A. Aldarf); bia@cs.ifmo.ru (I. A. Bessmertny)
https://github.com/Alaa-Shaker/ (A. Shaker);
https://itmo.ru/ru/viewperson/1134/bessmertnyy_igor_aleksandrovich.htm (I. A. Bessmertny)
0000-0003-2709-0766 (A. Shaker); 0000-0001-7249-5071 (A. Aldarf); 0000-0001-6711-6399 (I. A. Bessmertny)

humans can easily perform this process, but computers cannot. We need to understand how human builds an understanding of the text. Meaning is not a real or positive quality of a given word. It is a net of relations constructed in the text whose value is progressively determined during the reading process.

Furthermore, reading is not a neutral operation: to read is to determine meaning [1]. Recently, much research was dedicated to the quantum mechanical framework's application for information retrieval and natural language analysis. Van Rijsbergen, in 2004 was the first to propose using the quantum semantic model for information retrieval; he aimed to unify vector, logic, and statistical model [2]. All quantum-like models need to work in a formalized semantic space [1]. Hyperspace analog to language (HAL) model [3] is a widely used technique to formalize a textual space by representing the text as a square matrix. Platonov and Bessmertny, in their work [4] used the quantum semantic model for information retrieval in Russian and English languages. In this work, we will explore the potential of using a quantum-based model in information retrieval and its ability to analyze the Arabic natural language. Particularly, the model will use the HAL matrix to build a textual space, and then the Bell test will be used to detect whether the document is relevant to the user's query.

2. Main Work

In this paper, we try to apply a quantum theory to search in Arabic texts. Firstly, we need to form textual space, then the vector representation of text and pairs of words, after that basis, to study if there is quantum entanglement between two query words in the text using the Bell's test.

2.1. Hyperspace analog language

To apply quantum laws to find the relationship between two words in texts, we need to build a textual space. For that, we apply the HAL model [2], which is a widely used technique to formalize a document as a square matrix [5]. Also, HAL is a sensitive array for the links between two terms: row and column vectors record the co-occurrence information of previous and later words separately. The model is sensitive to direction, which helps to get the right representing of the context of the text that Boolean logic cannot do. For example, in the two statements, “Marx criticized the economists” and “The economists criticized Marx”. The queries “Marx”, “criticized” and “the economists” are commutative, whereas semantics seems not commutative [6, 7].

HAL vectors' value is affected by the window's size; a wider window means a greater chance for associations between two terms, but the large size may suffer from underfitting. On the other hand, the small window size means a strong association between two words and may suffer from overfitting. The words in the documents correspond to the labels of columns and rows. Each component of the word vector is inversely proportional to the distance between the considered word and the document's other corresponding words. We can limit the full context we are interested in by managing the window size. By setting the window's size, we can expand or reduce the width of the considered context.

Table 1
HAL matrix of the sentence "The basic concept of the word association"

	The	Basic	Concept	Of	Word	Association
The	2	3	4	5	0	0
Basic	5	0	0	0	0	0
Concept	4	5	0	0	0	0
Of	3	4	5	0	0	0
Word	5+1	2	3	4	0	5
Association	4	1	1	3	5	0

There is an example of the HAL matrix that can explain it clearly [8]. In this example, the sentence “The basic concept of the word association “with the window’s length W = 5 was used. The resulting matrix appears in Table 1. Each time the word pair occurs in the text the HAL matrix cell increments by the value W − D, where D − the distance between these words in the window.

In the Table 1, the intersection between the column labeled “Concept” and the row labeled “of” is equal to five because there is no word between them(5-0=5). This value decreases when the distance increases; for example, the intersection between the row labeled “the” and the column labeled “Basic” is equal to three because there are two words between them (5-2=3). Another example, we notice that intersection between the row labeled “word” and the column labeled “the” equals to (5 + 1), that because the word “the” occurs twice on the left of the word “word”, in the first occurrence there is no word between them so (5-0=5), and in the second one there are four words so (5-4=1); therefore the total equals to ((5-0)+(5-4)=5+1). We can see the HAL matrix’s direction sensitive feature.

2.2. Building Textual Space

As mentioned previously, we need to build a textual space that depends on the HAL square matrix, so we need to describe this space in geometric terms and provide it with the basis that generates its vectors. In the N-dimensional HAL matrix, each document will have an associated vector. The vector state of the document is the sum of all the words’ vectors the document contains. Each word vector state is extracted from the lines of the symmetric HAL matrix. The document vector state is defined as in equitation 1:

$$|\Psi\!> = \sum_{i}^{N} |W_i > \tag{1}$$

where $|\Psi >$ represents the vector of the whole document, and $|W_i >$ the vector of the word i.

We are now interested in analyzing how two words are connected within a document, namely, word A and word B. By concatenating between the corresponding row and column vectors from the HAL matrix related to the word, the two words can be represented as $\{|W_A >, |W_B >\}$.

By applying the Gram-Schmidt orthogonalization process to the non-orthogonal basis $\{|W_A >, |W_B >\}$ and $\{|W_B >, |W_A >\}$ getting two orthogonal coordinates: first orthogonal basis

$\{|u_A>, |u_{A\perp}>\}$ for the first word and $\{|u_B>, |u_{B\perp}>\}$ for the second one. Now we can project the document vector on these orthogonal bases, and the vector of the whole document can be represented as the following equitation:

$$\begin{aligned} |\Psi> &= a|u_A> +b|u_{A\perp}> \\ |\Psi> &= c|u_B> +d|u_{B\perp}> \end{aligned} \tag{2}$$

where the coefficients a, b, c and d of basis vectors can be calculated by projecting the vector of the document onto the bases vectors, for example, the coefficient (a) of vector $|u_A>$ can be calculated as in the following equitation:

$$a = \frac{<u_A|\Psi>}{\sqrt{<u_A|\Psi>^2 + <u_{A\perp}|\Psi>^2}} \tag{3}$$

2.3. Test Bell inequality (CHSH)

Bell inequality test is used in physics to determine whether there is entanglement between two particles. In this work, we investigate the ability to study the relation between two words in the text. This approach was proposed in [1, 9]. In this model, the following form of Bell's inequality CHSH is used as presented in the following equitation:

$$S_{bell} = |E(A,B) - E(A,C)| + |E(B,D) + E(C,D)| \tag{4}$$

Where A, B, C, and D are tests, and E (X, Y) stands for the expectation value of the outcome of mutual tests X and Y. The case where Bell's results in [2,$2\sqrt{2}$] can be achieved where the two particles under study are in entanglement. The value $2\sqrt{2}$, also known as the Tsirelsons bound [10, 11]. The zone between $2\sqrt{2}$ and 4 is called the "no-signaling" region. The maximum value is four can be attained with logical probabilistic constructions, often named Popescu and Rohrlich boxes (PR boxes) [12]. An area where the results are less than two that means there is no entanglement between the two particles.

2.3.1. The Bell's Inequality Operators:

The purpose of these operators is to quantify a query within our formalism. The query operators return $+1$ if the document meaning corresponds to the meaning we are interested in and -1 in the orthogonal direction. Operators spin Pauli matrix will be used, which is explicated as in equitation:

$$\hat{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{5}$$

The other operators can be defined using another Pauli matrix operator:

$$\hat{A}_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{6}$$

This operator switches the vector state components, which can be interpreted as a measure of a different meaning in the document.

More convenient with calculations to make all operations based on one basis, that is the basis of the word A shown in equitation.2 $\{|u_A>, |u_{A\perp}>\}$, for doing that we need to transform the operators $(\hat{B}, \hat{B_x})$ of basis $\{|u_B>, |u_{B\perp}>\}$ to the basis of the word A $\{|u_A>, |u_{A\perp}>\}$, to do that we need a transformation matrix from $\{|u_B>, |u_{B\perp}>\}$ to $\{|u_A>, |u_{A\perp}>\}$, which is shown in the following equitation:

$$M = \begin{pmatrix} <u_B|u_A> & <u_B|u_{A\perp}> \\ <u_{B\perp}|u_A> & <u_{B\perp}|u_{A\perp}> \end{pmatrix} = \begin{pmatrix} p & \sqrt{1-p^2} \\ -\sqrt{1-p^2} & p \end{pmatrix} \tag{7}$$

Now the operators of the basis word B can be transformed into basis associated with the word A by using the transformation matrix M. For example, in equitation 8, the operator $\hat{B}$ is converted to the basis of word A:

$$\hat{B} = M^{-1}.\hat{A}.M = \begin{pmatrix} 2.p^2-1 & 2.p.\sqrt{1-p^2} \\ 2.\sqrt{1-p^2} & 1-2.p^2 \end{pmatrix} \tag{8}$$

2.3.2. Form Bell's Inequality:

To determine the degree to which the document corresponds to the word A and word B simultaneously [13], a combination of operators($\hat{A}$,$\hat{B}$) with considering of using the Born rule can be used as shown in equitation (9):

$$<\hat{A}.\hat{B}>_\psi = <\psi|\hat{A}.\hat{B}|\psi> \tag{9}$$

Concretely, we calculate quantum means defined in equitation. 5, using different query operators which can be considered as measuring devices, and then determine the Bell query parameter:

$$S_{query} = \Big| <\hat{A}.\hat{B}_+>_\psi + <\hat{A}_x.\hat{B}_+>_\psi \Big| + \Big| <\hat{A}.\hat{B}_->_\psi - <\hat{A}_x.\hat{B}_->_\psi \Big| \tag{10}$$

where the other operators are defined as the following equitation:

$$\hat{B}_+ = -\frac{\hat{B}+\hat{B}_x}{\sqrt{2}}, \hat{B}_- = \frac{\hat{B}-\hat{B}_x}{\sqrt{2}}, \hat{B}_x = M^{-1}.\hat{A}_x.M \tag{11}$$

3. Results of Experiment

This section provides the results of this approach. The texts were taken from web sites in the Arabic language as a result of a Google search using a two-word query. The experiments of using a quantum-like semantic model for retrieving data in the Arabic language are divided into three sections.

In the first section, a quantum-like semantic model will be checked to find the entanglement of two words in the text. Our model's first step is to prepare the Arabic text for removing undesirable stop words. The second step is to use a stemmer to get the word's root because the same word takes many shapes depending on the grammatical construction. After that, we build an index of words included in the text to help in the HAL matrix building with different window

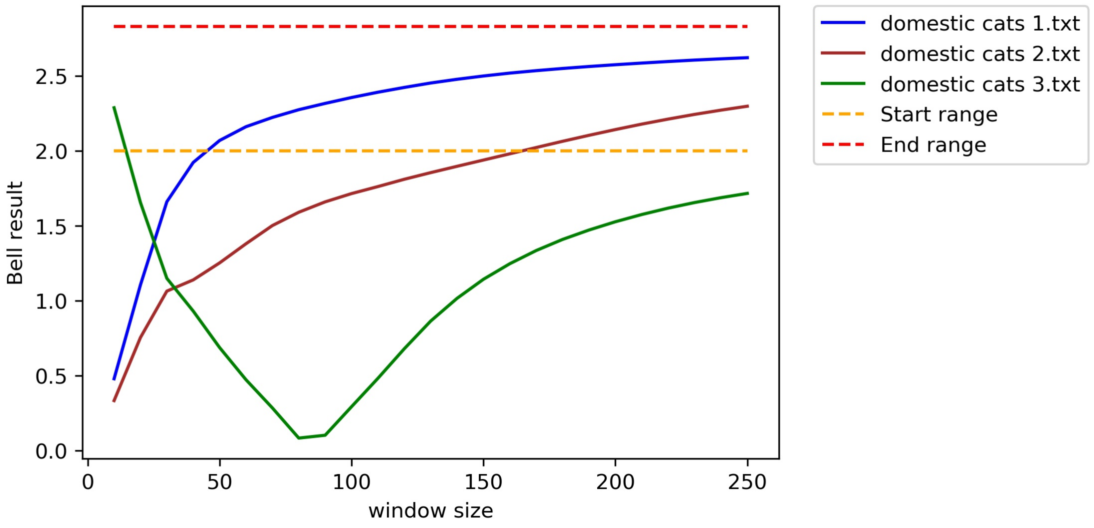

Figure 1: The results of Bell's inequality test on two extracted texts from Google search result.

sizes, in which each row and column corresponds to a unique word in the text, afterward, we extract the document vector and two vectors representing the query words. Next, a series of Bell's test measurements were performed for each window size used to build the HAL matrix.

The result of the performed experiment appears in Figure 1. The three texts were taken from Google's result. The blue line represents the result of Bell's test of the text entitled "domestic cats 1" in the Arabic language (1) related strongly to the search topic "domestic cats" in Arabic ( ). It is noted that from a suitable window size (50 words), the line reaches the beginning of the entanglement range, and with the increase of the HAL window size, an increase is observed with the values of the Bell's test results. This indicates a proportional dependency between the window size and the Bell's result. At the same time, a red line represents the result of Bell's test of text entitled "domestic cats 2", which is a little bit relevant to the subject search. We can note that this line does not reach the entanglement range until a considerable value (more than 150 words of HAL size window).

The green line represents the result of Bell's test; this text talks about specific news domestic cats' in New Zealand. This text is not closely related to the topic search. Therefore, this line does not reach the entanglement range.

In Figure 2, another example shows a series of Bell tests of three texts; the topic search was "Euphrates River" in the Arabic language ( ). It is noted that the line of text is entitled "Euphrates River 1", and the line of text is entitled "Euphrates River 2", these two lines reach the entanglement range because the two texts are related extensively to the topic search. During the previous two texts, a lot of information was mentioned about the river (riverbed, the civilizations based on it, its economic importance).

Another line entitled "Euphrates River 3" does not reach the threshold of entanglement because this text talks about a crisis that was happening to the river, so the text just mentioned the river's name as a location.

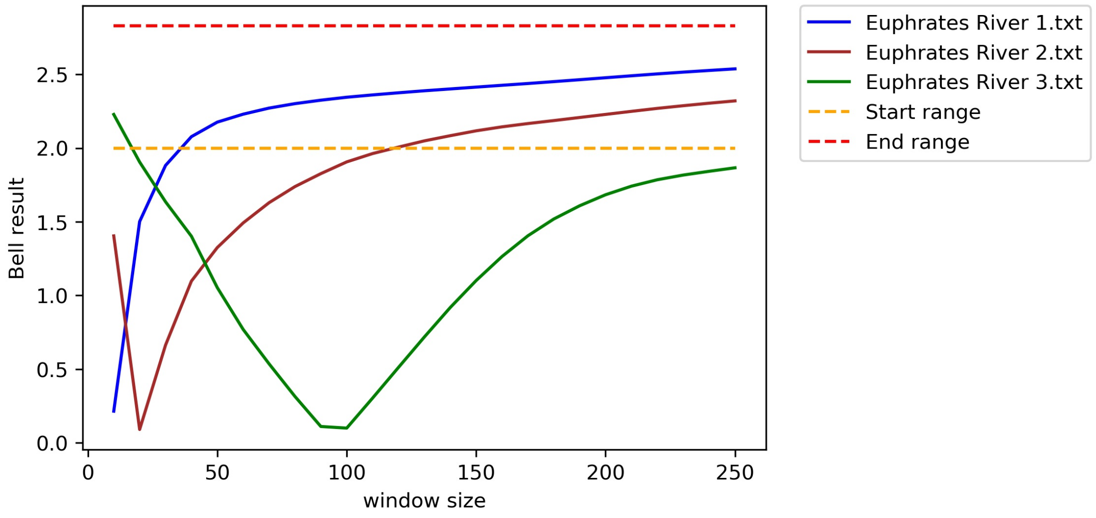

Figure 2: The results of Bell's inequality test on three extracted texts from Google search result.

In the second section, a quantum-like semantic model will be checked whether it can find the entanglement of two words in a large dataset of Arabic texts. For doing this mission, a dataset with five-hundred Arabic texts was collected. It contains six columns; each record includes fields for text, pair of words, and class to which type text belongs.

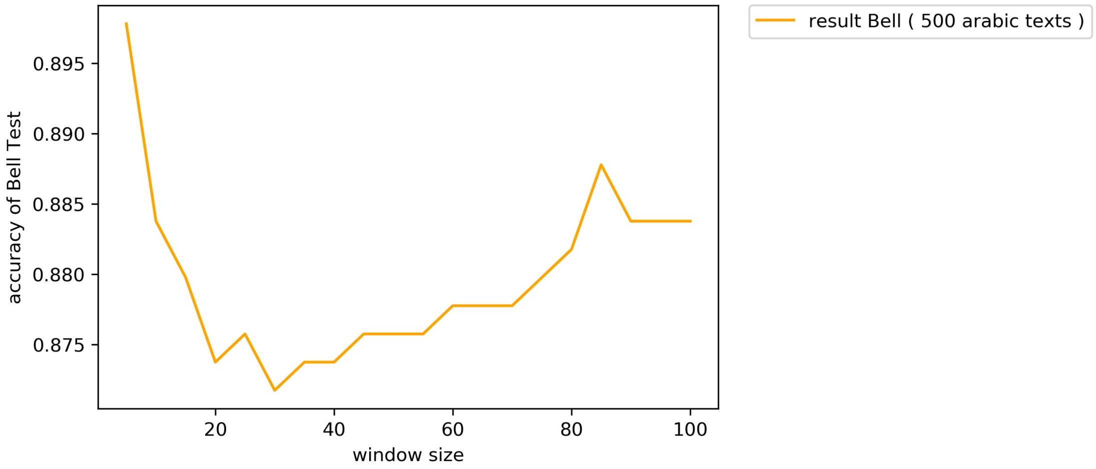

Figure 3: The precise of Bell test result with using five hundred Arabic texts.

Each text in the documents extracted from web pages is related to a pair of words. As mentioned above, each text will go through a set of processing stages that aims to delete stop words and symbols and obtain roots of the text's words. After that, we build a HAL matrix with different window sizes, vectorize text by getting three vectors for representing document and pair of words, then compute a value of Bell's test for each size of HAL window, to check the accuracy of each size of them.

Figure 3 shows the quantum semantic model's precision results in determining the degree of correlation between the two words of the title and the text.

It is noted that for a small window size of 5 words, the precision is 89.5%, but in the window's small size, the number of words studied is small, which means falling in under-fitting. On other side, a large window size, such as 100 words with precision 88.5% of detection relevance a text to the title, will lead to many words included in calculation operation, consequently, to over-fitting.

Under-fitting and over-fitting are considered as challenges facing the model and its ability to be generalized to new data. For this reason, the medium window size is chosen to avoid this problem, for example, 80 words, which corresponds to an accuracy of 89% in identifying whether the text belongs to the title. These results allow to conclude that the quantum-semantic model can be used to classify a textual dataset into pre-defined classes based on topic mining, an unsupervised text analytics algorithm used to find the group of texts related to the meaning.

The third section of the experiment is to test the quantum semantic model's ability to retrieve text. For this goal, an Arabic text dataset was built contains 100 Arabic texts related to ten different subjects (Syrian War, Ibn Sina (Avicenna), relativity theory, children's education, Pharaonic Civilization, Organic Chemistry, Euphrates river, Russian economy, Astronomy science, Psychology science). Each record contains five fields (the first one has Arabic text, the second and third fields contain pair of words related to the text, the fourth one for class, and the last field for saving source link).

To evaluate the quantum semantic model's performance, we use precision and recall, which are commonplace measures in information retrieval [14]. Another criterion for measuring the quality of an information retrieval system is the F-measure that combines precision and recall. The F-measure only produces a high result when Precision and Recall are both balanced.

Table 2 shows the values of the three criteria for evaluating the information retrieval system from executing ten queries.

Table 2
The results of the quantum-semantic model for ten query

Query	Precise	Recall	F-measure
Syrian war	0.43	0.90	0.581955
Son Sina	1.00	1.00	1.00
Relativity theory	0.91	1.00	0.95288
Children's education	0.625	1.00	0.769231
Pharaonic civilization	0.90	0.91	0.90
Organic chemistry	0.91	1.00	0.95288
Euphrates river	0.91	1.00	0.95288
Russian economy	0.669	1.00	0.80167
Astronomy science	0.455	1.00	0.61591
Psychology science	0.315	1.00	0.47908

The results listed in the Table 2 show the existence of discrepancy. We note that the precision values of four queries (Ibn Sina, relativity theory, organic chemistry, and the Euphrates River) were greater than 90 percent. Firstly, because these topics are explicit and the words used in them are in a “static” linguistic state, changes occur on them so limited, so it is easy to get the root for those words. For example (Ibn Sina, relativism, chemistry, river, the Euphrates) only exist in one linguistic case, which means that the two desirable words will be studied without neglecting them in some places or studying other words.

On the opposite side, we find a query such as “psychology science” that gets less accuracy because these words (psychology, science) are also used in texts classified as belonging to “Ibn Sina”, because he was a doctor and had some medical opinions in this field, and they also existed in the texts classified as belonging to children’s education.

4. Conclusion

The semantic vectors in HAL are representations for measures of context. The HAL method has already been used for the analogies with quantum theory by Bruza and Woods [15] and by Wittek and Dar´any [16] for extracting spectral content from the semantic space. HAL shows high potentiality because it is a simple way to build a semantic space. Bell’s test results in the first experiment, which appear in Figure 1 and Figure 2 show a strong correlation between Bell’s test and the window size used in building the HAL array. This correlation can be explained by the fact that with the increase in the HAL window size, more words will be involved in the study, and the length of the studied context will be longer. Therefore, we should choose the appropriate window size value, preferably medium. Fair window size value also helps to avoid study unnecessary words. The words that are located far from the query word have not got a relation with query words. Also, a small HAL window size means fewer context words will be studied, which may lead to ignoring the wanted words. In this experiment, we saw that we could get better ordering results related to the subject search.

The second experiment, which was conducted on a large number of texts in the Arabic language, shows the ability of the quantum semantic model to determine whether the studied text is related to the research topic or not. Furthermore, we can say that the quantum semantic model can be used to separate text into paragraphs based on specific topics, where the quantum semantic model determines whether the text or paragraph is relevant to the topic or not that depends on the result of Bell’s test, whether it is located in the range $[2, 2\sqrt{2}]$ or not, and its ability to classify texts into a pre-defined group.

From the testing of the model shown in this work, we see in the third experiment an attempt to retrieve texts related to the search’s subject from the database. According to the evaluation of more than one criterion, the results were acceptable to some extent, such as precision, recall, and F-measure.

The proposed model is still at the development stage and needs some improvements to include studying the entanglement between more than one pair of words and the possibility of merging it with other algorithms to be used in classifying and retrieving texts to give more accurate results.

References

[1] F. Galofaro, Z. Toffano, B.-L. Doan, A quantum-based semiotic model for textual semantics, Kybernetes 47 (2018) 307–320. doi:10.1108/K-05-2017-0187.

[2] V. Rijsbergen, C. Joost, The geometry of information retrieval, Cambridge University Press, Cambridge, England, 2004.

[3] K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence, Behavior research methods, instruments, & computers 28 (1996) 203–208. doi:10.3758/BF03204766.

[4] A. Trukhanov, A. Platonov, I. Bessmertny, Using quantum probability for word embedding problem, in: The 11th Majorov International Conference on Software Engineering and Computer Systems, volume 2590, CEUR Workshop Proceedings, ITMO University, Saint Petersburg, 2019.

[5] J. Barros, Z. Toffano, Y. Meguebli, B.-L. Doan, Contextual query using bell tests, in: International Symposium on Quantum Interaction, volume 8369, Springer, Berlin, Heidelberg, 2014, pp. 110–121. doi:10.1007/978-3-642-54943-4_10.

[6] F. Galofaro, Z. Toffano, B.-L. Doan, Linguistics and quantum theory: epistemological perspectives, in: 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), IEEE, Paris, 2016, pp. 660–668. doi:10.1109/CSE-EUC-DCABES.2016.257.

[7] D. Kartsaklis, Compositional operators in distributional semantics, Springer Science Reviews 2 (2014) 161–177. doi:10.1007/s40362-014-0017-z.

[8] N. Nasharuddin, Hyperspace analogue to language (hal): an example, 2012. URL: https://researchinbox.wordpress.com/2012/10/09/hal-example/.

[9] I. Bessmertny, X. Huang, C. Yu, A. Platonov, J. Koroleva, Applying the bell's test to chinese texts, Entropy 22 (2020) 275. doi:10.3390/e22030275.

[10] A. Cabello, Violating bell's inequality beyond cirel'son's bound, Physical review letters 88 (2002) 060403. doi:10.1103/PhysRevLett.88.060403.

[11] B. Cirel'son, Quantum generalizations of bell's inequality, Letters in Mathematical Physics 4 (1980) 93–100. doi:10.1007/BF00417500.

[12] S. Popescu, D. Rohrlich, Quantum nonlocality as an axiom, Foundations of Physics 24 (1994) 379–385. doi:10.1007/BF02058098.

[13] A. V. Platonov, E. A. Poleschuk, I. A. Bessmertny, N. R. Gafurov, Using quantum mechanical framework for language modeling and information retrieval, in: 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), IEEE, Almaty, Kazakhstan, 2018, pp. 1–4. doi:10.1109/ICAICT.2018.8747051.

[14] J. Euzenat, Semantic precision and recall for ontology alignment evaluation, in: IJCAI, volume 7, 2007, pp. 348–353.

[15] P. Bruza, J. Woods, Quantum collapse in semantic space: interpreting natural language argumentation, in: Second Quantum Interaction Symposium, Oxford University, 2008, pp. 141–147.

[16] P. Wittek, S. Dar ́anyi, Spectral composition of semantic spaces, in: D. Song, M. Melucci, I. Frommholz, P. Zhang, L. Wang, S. Arafat (Eds.), International Symposium on Quantum

Interaction, volume 7052, Springer, Berlin, Heidelberg, 2011, pp. 60–70. doi:10.1007/978-3-642-24971-6_7.

A. Online Resources

Here there are the links to texts used in the first experiment. The link to the dataset contains five-hundred Arabic texts used in the second experiment and the link to the dataset used in the third experiment to evaluate the result of queries.

- The source link of text entitled "domestic cats",
- The source link of text entitled "domestic cats 2",
- The source link of text entitled "domestic cats 3",
- The source link of text entitled "information engineering 1",
- The source link of text entitled "information engineering 2",
- The source link of text entitled "information engineering 3",
- The link of dataset used in the second experiment,
- The link of dataset used in the thrid experiment