WHY SEMANTIC SPACE IS 300-DIMENSIONAL: A POSSIBLE EXPLANATION

Semantic spaces (see, e.g., [2] and references therein) is a way to analyze the meaning of words from natural language. In this approach, we start with a large amount of texts in a given language – e.g., all the corresponding Wikipedia articles. For each pair of words x and y, we can gauge how close these two words are by comparing the number of times \( N(x, y) \) that both words occur in the same document in comparison with the numbers \( N(x) \) and \( N(y) \) of numbers of documents that contain, correspondingly, x and y. Based on 100 000 words, we get a 100 000 x 100 000 dimensional matrix. We then use known methods of dimension reduction to come up with a lower-dimensional matrix that captures all this information. In most cases, for different language, the resulting lower-dimensional semantic space is 300-dimensional. The question is: why?

Probably, one can explain this fact by some biological reasons, but this would leave us with another question: why is the human brain designed this way, that it uses a 300-dimensional space of meanings? From this viewpoint, the ultimate answer to our question has to come from the real physical world – what are the properties of the physical world that make the 300-dimensional space of meanings reasonable?

To answer this question, let us recall the main facts from physics; see, e.g., [1, 3]. According to modern physics, the world consists of elementary particles, and there are about \( 10^{90} \) of these particles. In this sense, to get a full description of the world, we need to store information about all these particles.

How many different properties do we need for this task? The simplest case is when we consider only binary properties, i.e., properties which can only be true or false. If we have one such property, we can describe 2 possible objects – one corresponding to true and one corresponding to false. If we have two such properties, then each of the previous cases splits into two new cases: corresponding to the values true or false of the second property. As a result, overall, we have \( 2 \times 2 = 2^2 = 4 \) objects. By using 3 properties, we can represent \( 2 \times 2^2 = 2^3 = 8 \) objects. In general, by using \( p \) properties, we can represent \( 2^p \) different objects.
How many properties do we need to represent all $10^{90}$ particles this way? By equating $2^p$ and $10^{90}$, we conclude that $2^p \sim 10^{90}$. We know that $2^{10} = 1024 \sim 10^3$. Thus, $10^{90} = (10^3)^{30} \sim (2^{10})^{30} = 2^{300}$. So, to describe the state of the world, we need $p = 300$ dimensions – exactly what is observed in semantic spaces.

Thus, physics indeed explains the 300-dimensional character of the semantic space.

References