Analysis of Geometric Arrangements of Distributed Representation of Words and Internal Structure of Word Co-occurrence Matrix

2130017 Akihiro Maeda

The evolution of natural language processing has been accelerating for the last decade. Large language models such as BERT and GPT-3 have shown their capacities to understand the language and to generate sentences humans can understand. The technological basis common to those models is the distributed representation of words, also called a word vector, which represents the meaning of words as a vector of several hundred dimensions.

It is well known that a word vector like word2vec is capable of solving analogy tests by vector arithmetic, implying that semantic relations are embedded in the vector space and are represented geometrically as a parallelogram.

The idea that the meaning of a word is distributively represented has been actively studied in psychology and cognitive science since the 1960s, and its linguistic theoretical basis can be traced back to Zellig Harris and John R. Firth's distributional hypothesis in the 1950s. Though the phenomena indicate that the latent mathematical structure of natural language may appear in the distributions of words as Harris claimed, it has yet to be clarified what structure the word vectors represent and how and why they work as in the large language models.

Motivated by Harris' structural linguistic view, this study aims to (1) clarify the distributional structure and semantic properties manifested in the distributed representation of words, (2) mathematically formulate the latent structure of the word co-occurrence matrix which is used to learn word vectors, and (3) demonstrate the existence of distributional structures in real corpora.

While a variety of word vectors have been proposed, they are largely divided into either the count type or the prediction type in terms of generation methods, both of which are learned from word co-occurrence statistics in corpora and encode similar semantic relations of words. The advanced studies on the word vectors suggest that distributional regularities, inherent in the language structure, already appear in a word co-occurrence matrix, which is to be used as an input to generate word vectors.

It is also shown that word vectors encode various types of semantic relations not limited to similarity, and that word vectors seem to have the linearity property in its vector space, though an exhaustive and systematic study of their mathematical structure and properties has yet to be done.

To elucidate the properties that word vectors obtain, first, by focusing on the analogical relation where word vectors form a parallelogram in the

vector space, we took a constructive approach and proposed a toy model where an artificial corpus replicates word vectors forming geometric shape such as parallelepiped. The model mathematically formulates the process by which the geometric relationship between distributed representations emerges from the corpus and allows algebraic analysis of internal structure of the co-occurrence matrix.

We mathematically derive the necessary and sufficient conditions for word vectors to form a parallelepiped in high-dimensional vector space. These conditions imply that a set of word vectors must be in the subspace of a certain kernel, which can be interpreted as the latent structure in the word distribution. It is revealed that the existence of certain linguistic relations, syntagmatic and paradigmatic, provide the necessary conditions of a parallelogram and that those relations can be described mathematically as complete bipartite subgraphs.

The proposed model is also extended beyond a parallelogram to cover other geometrical shapes together with corresponding toy corpus, suggesting the existence of more variety of semantic relations other than analogical ones.

Next, we empirically prove that the word vectors in the semantic relations are, in fact, geometrically arranged in high-dimensional vector spaces by using a real corpus.

We define a new metric which we call the geometric distance to the kernel to quantitatively evaluate how close a set of word vectors is to a parallelogram in space. We also proposed a method for normalizing the distance so that congruent and similar figures can be considered identical.

Using word vectors generated from the English Wikipedia dump corpus, we show that there is a significant difference between the distribution of geometric distances between a set of words in analogy relations of the Google analogy test set and randomly selected words. Furthermore, the defined distance reveals the relative positioning of word vectors among words in analogy relation, suggesting the difference of types of semantic relationship in the same category of the analogy test can be quantitatively identified by the proposed metric.

Though the result of the empirical test is significant, the accuracy to identify the analogy relation is still insufficient because the proposed normalization method resulted in too much variance of the measured distance, which suggests that the space where word vectors live may not be an Euclidean space.

Finally we propose a mathematical formulation to encode a general sentence into matrix representation and show that the distribution structure of a corpus can be represented as a tensor product. It is also shown that the word co-occurrence matrix can be derived from the tensor product by the

tensor network operation called mode-n unfolding.

This suggests the regularity and multi-linearity of the language structure and that its properties may be preserved in derived co-occurrence matrix and vector space. The decomposition of the tensor network would be the potential tool to extract the internal structure of the language distribution.

The contribution of this study is the mathematical formulation to algebraicallytreat the process of generating the word co-occurrence matrix from sentences in a corpus and to characterize the obtained word vectors geometrically in the vector space. It is an important insight that the distribution structure of a language can be a high-order tensor product, which suggests that observed regularities in word distribution derive from internal structures which might be obtained by tensor network decomposition. The proposed geometric distance is also promising because it can be a new research tool that enables us to locate the relative positional relationship between three or more words as a geometric property.

The remaining challenges and potentials for future research are as follows. First, the mathematical framework of the tensor network needs to be extended to formalize a complex system, including the recursive generation and hierarchical tree structure of the language.

Second, the fundamental question is why the algebraic structure emerges in a language system. The key finding is that paradigmatic relation can be defined mathematically in the bipartite graph structure.

Third, as an empirical issue, the space of word distributed representations may not be topologically homogeneous as in Euclidean space, suggesting that non-Euclidean spaces such as projective, hyperbolic, and others need to be studied.

Fourth, from the viewpoint of cognitive science, word vector operations may perform cognitive processes such as inference, as seen in large-scale language models such as BERT, which may reveal the computational model of the language.

This study paved the way for future research to construct a computational model of natural language based on word distributed representations and continued research is expected to be promising.