GALAXY INTERNATIONAL INTERDISCIPLINARY RESEARCH JOURNAL (GIIRJ)
ISSN (E): 2347-6915
Vol. 9, Issue 12, Dec. (2021)

DISTRIBUTIONAL SEMANTICS ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

Pulatova Niso
Termez State University, Foreign Philology Faculty
nisofo01@gmail.com, +998905191920

ANNOTATION

This article illustrates the main features of Distributional Semantics which is very influential semantic frameworks in Computational Linguistics and its importance in teaching a foreign language.

Keywords: Corpus linguistic, Computational Linguistics, Semantics, Distributional Semantics

INTRODUCTION

Distributive semantics is a field of linguistics that deals with the calculation of the degree of semantic proximity between linguistic units based on their distribution (distribution) in large arrays of linguistic data (text corpora). Each word is assigned its own context vector. The set of vectors forms a verbal vector space. The semantic distance between concepts expressed in natural language words is usually calculated as the cosine distance between the vectors of the word space. Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests include XML data management, query processing and optimization, bioinformatics, high dimensional indexing, parallel database systems, and cloud data management. He has published more than 100 research papers. Muter vision and transfer learning Distributive semantics is a field of linguistics that deals with the calculation of the degree of semantic proximity between linguistic units based on their distribution (distribution) in large arrays of linguistic data (text corpora). Each word is assigned its own context vector. The set of vectors forms a verbal vector space. The semantic distance between concepts expressed in natural language words is usually calculated as the cosine distance between the vectors of the word space.

"Distributional analysis is a method of language research based on the study of the environment (distribution, distribution) of individual units in the text and does not use information about the full lexical or grammatical meaning of these units" Within the framework of this method, an ordered set of universal procedures is applied to the texts of the studied language, which makes it possible to identify the main units of the language (phonemes, morphemes, words, word combinations), to classify them and to establish compatibility relations between them. Classification is based on the principle of substitution: language units belong to the same class if they can appear in the same contexts.

Distributional analysis was proposed by L. Bloomfield in the 20s of the XX century and was used mainly in phonology and morphology. Harris and other representatives of descriptive linguistics developed this method in their works in the 30-50s of the XX century. Similar ideas were put forward by the founders of structural linguistics, de Saussure and L. Wittgenstein. The idea of context vectors was proposed by psychologist Ch. Osgood in the framework of works on the representation of the meanings of words The contexts in which the words occurred acted

as dimensions of multi-bit vectors. As such contexts, Osgood's works used antonymic pairs of adjectives (for example, fast-slow), for which the survey participants scored on a seven-point scale.

The term context vector was introduced by S. Gallant to describe the meaning of words and to resolve lexical ambiguity.Gallant's work used a variety of attributes specified by the researcher, such as a person, a man, a car, etc. An example of a context feature space describing the meaning of the word astronomer from Gallant's work.

Over the past two decades, the method of distributional analysis has been widely applied to the study of semantics. We have developed a distributional-semantic methodology and corresponding software that allows us to automatically compare the contexts in which the studied language units meet and calculate the semantic distances between them

Psychological experiments have confirmed the truth of this hypothesis. For example, in one of the works, the participants of the experiment were asked to express their opinion about the synonymy of the pairs of words presented to them. The survey data was then compared with the contexts in which the words being studied occurred. The experiment showed a positive correlation between the semantic proximity of words and the similarity of the contexts in which they occur. Vector spaces from linear algebra are used as a way to represent the model. Information about the distribution of linguistic units is presented in the form of multi-bit vectors that form a verbal vector space. Vectors correspond to linguistic units (words or phrases), and dimensions correspond to contexts. The coordinates of the vectors are numbers that show how many times a given word or phrase has occurred in a given context.

MODELS OF DISTRIBUTIVE SEMANTICS

There are many different models of distributive semantics, which differ in the following parameters: context type: context size, right or left context, ranking; quantification of the frequency of occurrence of a word in this context: absolute frequency, TF-IDF, entropy, joint information, etc.; measure of the distance between vectors: cosine, scalar product, Murkowski distance, etc. ; method of reducing the dimension of the matrix: random projection, singular value decomposition, random indexing, etc. The following distributional-semantic models are most widely known: Model of vector spaces Latent semantic analysis Thematic modeling Predictive models .Reducing the dimension of vector spaces.

When applying distributional-semantic models in real applications, the problem arises of too large dimension of vectors corresponding to a huge number of contexts represented in the text corpus. There is a need to use special methods that allow you to reduce the dimension and sparsity of the vector space and at the same time save as much information as possible from the original vector space. The resulting compressed vector representations of words in English terminology are called word embedding's. Methods for reducing the dimension of vector spaces: deleting certain vector dimensions according to linguistic or statistical criteria; singular value decomposition; principal Component Analysis (PCA); random indexing. Predictive Models of Distributive Semantics[edit] Another way to obtain small — dimensional vectors is machine learning, in particular artificial neural networks. When training such predictive models, the target representation of each word is also a compressed vector of relatively small size (English

embedding), for which, during multiple passes through the training corpus, the similarity with the vectors of neighbors is maximized and the similarity with the vectors of words that are not its neighbors is minimized.However, unlike traditional counting models (eng . count models), in this approach, there is no stage of reducing the dimension of the vector, since the model is initially initialized with vectors of small dimension (on the order of several hundred components). Such predictive models represent the semantics of natural language more accurately than counting models that do not use machine learning. The most well — known representatives of this approach are the Continuous Bag-of-Words (CBOW) and Continuous Skipgram algorithms, first implemented in the word2vec utility, introduced in 2013.An example of applying such models to the Russian language is presented on the RusvectōRēs web service.

Application areas edit Models of distributive semantics have found application in research and practical implementations related to semantic models of natural language. Distribution models are used to solve the following problems. Identification of semantic proximity of words and phrases automatic clustering of words by the degree of their semantic proximity; automatic generation of thesauri and bilingual dictionaries resolving lexical ambiguity; expanding queries by using associative links; defining the subject of the document; clustering of documents for information search; extracting knowledge from texts; construction of semantic maps of various subject areas; modeling peripherals; determining the tone of an utterance; modeling of combinability constraints of words.

A lot of work undergone over this field Shuang Li received the Ph.D. degree in control science and engineering from the Department of Automation, Tsinghua University, Beijing, China, in 2018. He was a Visiting Research Scholar with the Department of Computer Science, Cornell University, Ithaca, NY, USA, from November 2015 to June 2016. He is currently an Assistant Professor with the school of Computer Science and Technology, Beijing Institute of Technology, Beijing. His main research interests include machine learning and deep learning, especially in transfer learning and domain adaptation

Binhui Xie is a graduate student at the School of Computer Science and Technology, Beijing Institution of Technology. His research interests focus on computer vision and transfer learning Bin Zang is a graduate student at the School of Computer Science and Technology, Beijing Institution of Technology. His research interests focus on Chi Harold Liu receives the Ph.D. degree from Imperial College, UK in 2010, and the B.Eng. degree from Tsinghua University, China in 2006. He is currently a Full Professor and Vice Dean at the School of Computer Science and Technology, Beijing Institute of Technology, China. Before moving to academia, he joined IBM Research - China as a staff researcher and project manager, after working as a postdoctoral researcher at Deutsche Telekom Laboratories, Germany, and a visiting scholar at IBM T. J. Watson Research Center, USA. His current research interests include the Big Data analytics, mobile computing, and deep learning. He has published more than 90 prestigious conference and journal papers and owned more than 14 EU/U.S./U.K./China patents. He is a Fellow of IET, and a Senior Member of IEEE.

Xinjing Cheng is the head of perception team at Inceptio Tech., Shanghai, China. Before that, he was a research assistant with the Intelligent Bionic Center, Shenzhen Institutes of

Advanced Technology (SIAT), Chinese Academy of Sciences(CAS), Shenzhen, China. His current research interests include computer vision, deep learning, robotics and autonomous driving.

Ruigang Yang received his Ph.D. degree from University of North Carolina at Chapel Hill and M.S degree from Columbia University. He is the CTO of Inceptio and a full professor of computer science at the University of Kentucky. He was the director of Robotics and Autonomous Driving Lab at Baidu Research. He has published over 130 papers, which, according to Google Scholar, has receive over 14000 citations with an H-index of 61. He has received a number of awards, including US NSF Career award in 2004 and the Deans Research Award at the University of Kentucky in 2013.

Guoren Wang received the BSc, MSc, and PhD degrees from the Department of Computer Science, Northeastern University, China, in 1988, 1991 and 1996, respectively. Currently, he is a Professor and the Dean with the School of Computer

Distributive semantics is a field of linguistics that deals with the calculation of the degree of semantic proximity between linguistic units based on their distribution (distribution) in large arrays of linguistic data (text corpora).In Uzbekistan some scholars are working over it. The semantic distance between concepts expressed in natural language words is usually calculated as the cosine distance between the vectors of the word space. In research interests include XML data management, query processing and optimization, parallel database systems, and cloud data management. In research involved such scholars like A.Po`lotov, S.Muhamedov, M.Ayimbetov, S.Muhamedova, S.Karimov, G.Jumanazarova, A.Babanarov, D.Yerboeva, N.Abdurahmonova, A.Norov they have published many research papers. We hope many works will be carried out in this field.

REFERENCES

1. Лингвистический энциклопедический словарь / Ярцева В. Н.. — М.: Советская энциклопедия, 1990.
2. Osgood C., Suci G., Tannenbaum P. The measurement of meaning (англ.). — University of Illinois Press, 1957.
3. Gallant S. Context vector representations for document retrieval (англ.) // Proceedings of AAAI Workshop on Natural Language Text Retrieval : конференция. — 1991.
4. Митрофанова О.А. Измерение семантических расстояний как проблема прикладной лингвистики (рус.) // Структурная и прикладная лингвистика. Межвузовский сборник : журнал. — Издательство СПбГУ, 2008. — Вып. 7.
5. Rubenstein H., Goodenough J. Contextual correlates of synonymy (англ.) // Communications of the ACM : журнал. — 1965. — Vol. 8, iss. 10. — P. 627—633.
6. Rubenstein H., Goode Sahlgren M. An Introduction to Random Indexing (англ.) // Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005 : конференция. — 2005. Архивировано 8 марта 2014 года.
7. Kutuzov A., Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian (англ.) // Сборник "Компьютерная лингвистика и

интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27 — 30 мая 2015 г.)" : конференция. — 2015. — Vol. 21, iss. 14.

8. Baroni, Marco and Dinu, Georgiana and Kruszewski, German. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. — 2014. — № 1. — С. 238—247.nough J. Contextual correlates of synonymy (англ.) // Communications of the ACM : журнал. — 1965. — Vol. 8, iss. 10. — P. 627—633.