Distributional Semantics

Distributional semantics develops theories and methods to represent the meaning of natural language expressions, with vectors encoding their statistical distribution in linguistic contexts. It is at once a theoretical model to express meaning, a practical methodology to construct semantic representations, a computational framework for acquiring meaning from language data, and a cognitive hypothesis about the role of language usage in shaping meaning. This book aims to build a common understanding of the theoretical and methodological foundations of distributional semantics. Beginning with its historical origins, the text exemplifies how the distributional approach is implemented in distributional semantic models. The main types of computational models, including modern deep neural language models, are described and evaluated, demonstrating how they address various types of semantic issues. Open problems and challenges are also analyzed. Students and researchers in natural language processing, artificial intelligence, and cognitive science will appreciate this book.

ALESSANDRO LENCI, Ph.D., is Professor of Linguistics and Director of the Computational Linguistics Laboratory at the University of Pisa. His research interests include distributional semantics, computational linguistics, and natural language processing. The recipient of the "10-year Test-of-Time-Award" from the Association for Computational Linguistics in 2020, he has published extensively and coordinated many projects on natural language processing and cognitive science.

MAGNUS SAHLGREN, Ph.D., is Head of Research for Natural Language Understanding at AI Sweden. Known primarily for research on computational models of meaning, Sahlgren's work lies at the intersection of computational linguistics, machine learning, and artificial intelligence. His doctoral dissertation, entitled "The Word-Space Model," was awarded the prize for the "Most Prominent Scholarly Achievement of 2006" by the Stockholm University Faculty of Humanities.

STUDIES IN NATURAL LANGUAGE PROCESSING

Volumes in the SNLP series provide comprehensive surveys of current research topics and applications in the field of natural language processing (NLP) that shed light on language technology, language cognition, language and society, and linguistics. The increased availability of language corpora and digital media, as well as advances in computer technology and data sciences, has led to important new findings in the field. Widespread applications include voice-activated interfaces, translation, search engine optimization, and affective computing. NLP also has applications in areas such as knowledge engineering, language learning, digital humanities, corpus linguistics, and textual analysis. These volumes will be of interest to researchers and graduate students working in NLP and other fields related to the processing of language and knowledge.

Also in the Series

Douglas E. Appelt, *Planning English Sentences*

Madeleine Bates and Ralph M Weischedel (eds.), *Challenges in Natural Language Processing*

Steven Bird, *Computational Phonology*

Peter Bosch and Rob van der Sandt, *Focus*

Pierette Bouillon and Federica Busa (eds.), *Inheritance, Defaults and the Lexicon*

Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Varile, Annie Zaenen, Antonio Zampolli, and Victor Zue (eds.), *Survey of the State of the Art in Human Language Technology*

David R. Dowty, Lauri Karttunen, and Arnold M Zwicky (eds.), *Natural Language Parsing*

Ralph Grishman, *Computational Linguistics*

Graeme Hirst, *Semantic Interpretation and the Resolution of Ambiguity*

András Kornai, *Extended Finite State Models of Language*

Kathleen R McKeown, *Text Generation*

Martha Stone Palmer, *Semantic Processing for Finite Domains*

Terry Patten, *Systemic Text Generation as Problem Solving*

Ehud Reiter and Robert Dale, *Building Natural Language Generation Systems*

Manny Rayner, David Carter, Pierette Bouillon, Vassilis Digalakis, and Matis Wiren (eds.), *The Spoken Language Translator*

Michael Rosner and Roderick Johnson (eds.), *Computational Lexical Semantics*

Richard Sproat, *A Computational Theory of Writing Systems*

George Anton Kiraz, *Computational Nonlinear Morphology*

Nicholas Asher and Alex Lascarides, *Logics of Conversation*

Margaret Masterman (edited by Yorick Wilks), *Language, Cohesion and Form*

Walter Daelemans and Antal van den Bosch, *Memory-based Language Processing*

Laurent Prévot (eds.), *Ontology and the Lexicon: A Natural Language Processing Perspective*

Thierry Poibeau and Aline Villavicencio (eds.), *Language, Cognition, and Computational Models*

Bind Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Second Edition*

Marcos Zampieri and Preslav Nakov, *Similar Languages, Varieties, and Dialects: A Computational Perspective*

Tommaso Caselli, Eduard Hovy, Martha Palmer, Piek Vossen, *Computational Analysis of Storylines: Making Sense of Events*

Piek Vossen and Antske Fokkens (eds.), *Creating a More Transparent Internet*

Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari,

Kiyong Lee, *Annotation-based Semantics for Space and Time in Language*

"Lenci and Sahlgren's textbook is a landmark contribution to the fast-growing and increasingly important discipline of distributional semantics. They have managed to distill 60 years of diverse research on distributional semantics, from its beginning in structural and corpus linguistics and psychology, through the application of techniques from information retrieval and linear algebra, to the most recent developments driven by deep neural networks and large language models in NLP. The authors synthesize the major findings from different fields and integrate these diverse traditions into a comprehensive and coherent framework of distributional meaning. Lenci and Sahlgren's text promises to be the new standard for reference and teaching in this area."

– James Pustejovsky, Brandeis University

Distributional Semantics

ALESSANDRO LENCI
University of Pisa

MAGNUS SAHLGREN
AI Sweden

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,

New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107004290

DOI: 10.1017/9780511783692

First published 2023

A catalogue record for this publication is available from the British Library

A Cataloging-in-Publication data record for this book is available from the Library of Congress

ISBN 978-1-107-00429-0 Hardback

Contents

Preface

What Is Distributional Semantics?

Language is used to convey *meaning*, and meaning is the quintessence of language. In fact, meaning is the "holy grail" (Jackendoff, 2002) of linguistics, philosophy, psychology, and neuroscience, as well as of Natural Language Processing (NLP) and Artificial Intelligence (AI). No research field that has at its center language and its use in communication by human or artificial agents can avoid inquiring into issues such as: What is the meaning of linguistic expressions at different levels of complexity (e.g., words, phrases, sentences, etc.)? How can meaning be represented? How is meaning structured and organized in language(s) and in the human mind? How does meaning represent the world? What meaning relations exist between words and sentences? How are word meanings combined to form the interpretation of sentences and discourses? How is meaning acquired? How does meaning interface with aspects of human cognition such as concepts, reasoning, perception, and action?

These and other similar questions are the object of study of *semantics*. This book is concerned with *distributional semantics*, an approach to meaning that develops theories and methods for quantifying and categorizing semantic properties of linguistic items based on their distributional properties in text corpora, as samples of language data. Grounded in the *Distributional Hypothesis*, according to which words with similar linguistic contexts tend to have similar meanings, the essence of distributional semantics can be described as follows:

> The meaning of a linguistic expression is represented with a real-valued *vector* (nowadays commonly called *embedding*) that encodes its statistical distribution in contexts.

The continuous nature of its representations distinguishes distributional semantics from other theoretical frameworks that instead represent meaning with symbolic structures. Moreover, there is a large array of computational methods to learn semantic representations from corpus data, which are then used in NLP and AI applications, as well as for cognitive and linguistic analyses. We refer to such methods as *distributional semantic models* (DSMs).

Distributional semantics lies at the crossroad of computational linguistics, AI, and cognitive science. Therefore, it is the combination of various characters: (i) it is a theoretical model to represent meaning, (ii) a computational framework to acquire it from language data, (iii) a practical methodology to construct semantic representations, and (iv) a cognitive hypothesis about the role of language usage in shaping meaning. This makes distributional semantics a fascinating research area and a privileged vantage point to combine theoretical, cognitive, and computational perspectives on language.

The Need for a Common Ground

Distributional semantics is a relatively novel approach to the study of meaning, but it has undergone deep transformations since its outset. Three main generations of models have followed one another: (i) *count DSMs* that build distributional vectors by recording co-occurrence frequencies; (ii) *predict DSMs* that learn vectors with shallow neural networks trained to predict surrounding words; (iii) *contextual DSMs* that use deep neural language models to generate inherently contextualized vectors for each word token, and therefore radically depart from previous *static DSMs* that instead learn a single vector per word type. Across its history, the changes in distributional semantics involve the way to characterize linguistic contexts, the methods to generate word vectors, the nature of such vectors, and the model complexity itself, which has exponentially grown especially with the last generation of deep neural DSMs.

Though it has been extensively practiced since the 1990s, the popularity of distributional semantics has exploded only recently, thereby becoming the leading approach to meaning representation in computational linguistics and cognitive science. This has coincided with the advent of neural network methods as the dominant computational paradigm for AI and NLP, which has produced a proliferation of models to learn semantic vectors and has fostered their use in downstream applications. Several "off-the-shelf" pretrained vectors are nowadays available, and provide ready-to-use resources for experimentation and system development. The fast and unprecedented expansion of distributional semantics is surely a positive fact that has boosted research in this area, but it also has its drawbacks. Like many novel scientific fields, distributional semantics lacks a common understanding of its theoretical and methodological foundations. This often prompts a rather opportunistic attitude, which not only fosters rapid development but also risks hindering progress by constantly "reinventing the wheel," and encourages a strong tendency to focus on the last "trendy" model, without considering whether and to what extent it brings significant advances over the previous ones, whether it can address qualitatively different issues or simply provides some quantitative improvement on old problems. Moreover, pretrained distributional vectors are often used in a kind of "black box" fashion, with little awareness of the effects of the different training settings and of the general properties and limits of distributional representations. This leads to a dangerous polarization between enthusiastic and skeptical attitudes toward distributional approaches to meaning, which prevents a rational and critical analysis of their actual potentialities.

Our goal with this book is to contribute to establishing distributional semantics as a *unitary and general framework to represent and study meaning*, which is implemented in various computational models and applied to address different semantic issues. We aim at building a common understanding of the theoretical and methodological foundations of distributional semantics by:

- tracing its historical origins;
- discussing and clarifying what is meant by *distributional* and *semantic* in this discipline;
- exemplifying how the distributional approach is implemented in DSMs;
- identifying, distinguishing, and evaluating the main types of DSMs;
- presenting how various types of semantic issues are addressed with DSMs;
- analyzing the open problems and challenges.

Distributional semantics is a fast and continuously moving field. It would be as impossible as it is useless to provide an exhaustive account of all its variations. Therefore, some of them have been left out by necessity. Up-to-date details may be easily obtained by browsing online resources such as the *ACL Anthology* or *arXiv*. This book has a different and more ambitious aim: Abstracting from the wealth of works in this area a general view that may help the reader to understand and master the common methodological paradigm that lies behind the increasingly large number of studies in distributional semantics.

Outline of the Book

The book consists of three main parts. *Part I – Theory* presents the major methodological aspects of distributional semantics as a framework for meaning investigation. Chapter 1 deals with the epistemological principles of distributional semantics, in particular the Distributional Hypothesis, its historical foundations, and its place in theoretical and computational linguistics. Chapter 2 introduces the basic notions to construct distributional semantic representations from corpora: various types of linguistic contexts, vectors, co-occurrence matrices and their dimensionality reduction to extract latent dimensions, vector similarity measures, and so on.

Part II – Models presents the most important families of static DSMs. Chapter 3 contains a synoptic view of the different types and generations of models. We then focus on static DSMs, since they are still the best-known and most widely studied family of models, and they learn context-independent distributional representations that are useful for several linguistic and cognitive tasks. Chapter 4 illustrates matrix models, which organize distributional information into co-occurrence matrices, Chapter 5 discusses models based on the use of random patterns to accumulate low-dimensional vectors in an incremental way, and Chapter 6 reviews predict DSMs that use shallow neural language models.

Part III – Practice explores how distributional semantics addresses different aspects of meaning. Chapter 7 focuses on the main methods and benchmarks to evaluate DSMs and analyzes the results of a comparative evaluation of the models described in Part II on a large range of tasks and datasets. Chapter 8 illustrates the most important applications of distributional semantics to the study of lexical meaning (e.g., word senses, paradigmatic relations, etc.). Chapter 9 discusses the main distributional approaches to the compositional interpretation of complex linguistic expressions and the last generation of deep neural language models that learn contextual embeddings and provide new tools to explore the context-sensitive nature of meaning.

Chapter 10 concludes the book with a general summary of the current state of distributional semantics, its future prospects, and challenges.

Distributional semantics is a multidisciplinary field. This book is intended for NLP and AI scholars and practitioners, linguists, and cognitive scientists. The background and interests of this potential audience are extremely heterogeneous, like the toolbox of distributional semantics, which includes notions from theoretical and computational linguistics, linear algebra, statistics, neural networks, and so on. We have tried to make this book as self-contained as possible, and we have added mathematical notes that introduce the key concepts. We expect the book to be accessible to a large number of readers from different fields, but of course it has been impossible to present all the mathematical and computational methods used by DSMs and their applications. However, gaps will be easily filled by consulting general handbooks or introductory texts, some of which have been listed in the further reading section at the end of each chapter.

Terminological Issues

Distributional semantics is mainly concerned with the *lexicon*, which includes *words* as well as complex *set phrases* and *constructions*, such as idioms and multiword expressions. In computational linguistics and information retrieval, words and set phrases are also referred to as *terms*. In this book, we use *lexeme*, *lexical item*, *term*, and *word* interchangeably, unless differently specified, because the concept of term has a strong overlap with the concept of lexeme, and words are the most typical targets of DSMs.

We distinguish between meaning and sense. We use *meaning* to refer to the general semantic content of a lexeme and *sense* to refer to a particular use of a lexeme in a context (or class of contexts). Theories of lexical meaning typically assume the notion of sense as their starting point. On the other hand, distributional semantics dispenses with senses as a primitive theoretical notion. In fact, distributional representations are primarily representations of the content (i.e., meaning) of a lexeme, rather than of its senses. Chapter 8 discusses how the notion of sense can be represented in DSMs.

The widespread adoption of neural networks in distributional semantics has popularized the term "embedding," which is, however, used in different ways, sometimes to mean a vector learnt by neural models, sometimes as a generic equivalent of distributional representation. In this book, we use *distributional vector* to refer to any vector that encodes distributional information. Instead, we reserve *embedding* for all kinds of low-dimensional, dense vectors, independently of the method that creates them.

Acknowledgments

Writing this book took much longer than we expected. However, this delay has also had positive effects because we have been able to witness the rapid and dramatic changes that distributional semantics has undergone in the last years.

We are very grateful to everyone at Cambridge University Press who helped us throughout this work, in particular Becca Grainger, Lauren Cawles, Johnathan Fuentes, and Kaitlin Leach. We would like to thank Chu-Ren Huang for having believed in this project since its outset and for having patiently, but trustfully, waited for its completion. The first stages of this work were carried out with Marco Baroni. Though life and work lead us along different roads, we cannot forget the inspiring discussions we had on these themes.

Our gratitude goes to Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani for their precious work on the experiments included in Chapter 7; Giovanna Marotta for insightful comments on Chapter 1; and Alessandro Bondielli, Lucia Passaro, and Paolo Pedinotti for having read and commented on the manuscript. We thank all our past and present students and collaborators, who have always been our compass in writing the book.

We deeply thank our beloved ones for the constant support they gave us throughout this long journey. Without their help, patience, and understanding, none of this would have been possible.