Natural Language Engineering (2023), **29**, pp. 1411–1413
doi:10.1017/S1351324923000074

CAMBRIDGE UNIVERSITY PRESS

Book Review

***Syntactic n-grams in Computational Linguistics*, by Grigori Sidorov. Cham, Springer Nature, 2019. ISBN 9783030147716. IX + 92 pages.**

Over the past decade, natural language engineering has highlighted the robustness of data-intensive scientific discovery (see e.g., Szalay and Gray 2006), ranging from linguistic datafication to computational modelling. Using quantitative approaches to formulate hypotheses and theories has become an important research trend in data-intensive disciplines that involve automatic language processing and language model construction, such as corpus linguistics and computational linguistics. In this respect, Grigori Sidorov's recent work entitled *Syntactic n-grams in Computational Linguistics* in the *Springer Briefs in Computer Science* series is an interesting publication contributing to such research areas.

The book provides a comprehensive analysis of how to construct syntactic n-grams (henceforth abbreviated as 'sn-grams') in a non-linear structure, in which lexical dependency is investigated from the angle of syntactic relations in syntactic trees. The approach presented differs from the traditional n-gram perspective that highlights the linear structure of texts, focusing more on the dependency relationship between words that are not necessarily 'neighboured'. Both continuous and non-continuous sn-grams are taken into consideration and are explained with real-world examples, utilising a number of novel ideas in terms of methodological innovation. This in-depth monograph introducing the basics and potentials of sn-grams would appeal to a wide spectrum of readership, such as computational linguists, corpus linguists, natural language engineers and general language researchers who wish to employ quantitative approaches in their studies.

The book features an appealing way of organising research themes throughout the text, falling under two major headings: 'Vector Space Model in the Analysis of Similarity between Texts' (Part I) and 'Non-linear Construction of n-grams' (Part II). Following a clear overview of the key issues to be addressed across different chapters, the book starts with an introduction to the fundamental knowledge of computational linguistics, by means of an extensive description of artificial intelligence (Chapter 1). In computational linguistics, large-scale language data such as corpora are employed to help construct machine-readable language models, incorporating computer-assisted language processing techniques and the rationale of statistics into the analysis of human languages. In this sense, what computational linguistics essentially covers is how to 'marry' computer science and linguistics, and how to construct/train machine learning models by selecting 'the features (and their values) that are introduced into the classification and clustering algorithms' (p. 4).

One of the most commonly used machine learning models is the vector space model (Chapter 2). The main idea of the vector space model indicates the comparison of objects in a formal way, either by using 'a representation with features (characteristics) and their values' or by using 'associative memories' (p. 5) between objects. Sidorov delineates the rationale of the vector space model with an example of comparing two books and provides a step-by-step instruction of how to select features and how to formally represent the vector spaces. With regard to similarity

https://doi.org/10.1017/S1351324923000074 Published online by Cambridge University Press

of objects in the vector space model, each object is an N dimension vector. The closer the object pair is in the dimension, the greater the similarity between the two objects, and vice versa.

Based on the rationale of the vector space model, Sidorov discusses the application of this model to text comparison and analysis (Chapter 3). He briefly reviews the procedure for measuring the similarity between texts with regard to common computational linguistics tasks, arguing that the easiest method for retrieving information is to utilise words as textual features. In addition, Sidorov proposes the use of n-grams as features in the vector space model and discusses the approach to calculate the values of n-grams.

Based on a brief review of the methods of investigating the dependencies between features, Chapter 4 deals with the latent semantic analysis that can be used for reducing dimensions in a vector space. Well-crafted examples are provided to give readers a clear picture of how to combine dimensions and reduce redundant features while keeping data changes to a minimum. As Sidorov notes, it is important to decide the number of dimensions in a new vector space, for which he proposes the values between 100 and 500 subject to the nature of tasks. He also suggests the use of existing R or Python codes to carry out the latent semantic analysis.

Chapter 5 looks primarily at the design of experiments in computational linguistics, in which nine major steps are outlined in detail, from task defining and text selection to textual data conversion and value calculation, while Chapter 6 offers a few examples of applying n-grams to authorship attribution with syllables. Sidorov proposes that syllables can be employed as features in the vector space model, something which has not been adopted in previous studies pertinent to authorship attribution. With the aid of automatic syllabification techniques, he categorises character n-grams as affix character n-grams, word character n-grams and punctuation character n-grams, and carries out two experiments based on the data retrieved from an English corpus and a self-built Spanish cross-topic corpus. The results of his research confirm the validity of character n-gram-based approach to a large extent. Part I ends with an in-depth discussion regarding the interplay between deep learning and the vector space model (Chapter 7).

Part II concentrates on the non-linear construction of sn-grams. In Chapter 8, a set of work related to the application of syntactic information is reviewed in terms of distributional semantics, selection preferences, information extraction, improvement of dependency parsing, and online resources of sn-grams. Sidorov proposes the construction of n-grams based on syntactic trees and examines methods of sn-gram segmentation in both Spanish and English. Readers are familiarised with various programming techniques using tools such as FreeLing and Stanford Parser for segmenting n-grams. Sidorov classifies sn-grams according to their components (Chapter 9); specifically, lexical elements, part-of-speech tags, tags of syntactic relations, characters and mixed sn-grams.

Chapter 10 demonstrates a distinction between continuous sn-grams and non-continuous sn-grams. As illustrated in syntactic trees, continuous sn-grams do not allow bifurcation, in which words that constitute sn-grams are considered as lexical wholes (e.g., formulaic sequences), while non-continuous sn-grams allow bifurcation, in which words making up sn-grams are semantically related but not necessarily formally connected or sequenced.

Chapter 11 deals with the method of representing non-continuous sn-grams in the metalanguage. Sidorov proposes a number of conventions by separating 'the continuous elements of n-grams with whitespaces', putting 'commas in the bifurcation parts' as well as using 'parentheses to mark the bifurcation parts in order to avoid the structural ambiguity' (p. 69). Two examples are offered to clarify the procedure for retrieving non-continuous sn-grams automatically (Chapter 12). The FreeLing parser and the Stanford parser are used for generating syntactic trees, respectively, in Spanish and English, while useful source codes are also presented to help motivate readers to write creative programmes related to n-gram processing.

Chapter 13 focuses on the application of sn-grams to authorship attribution. Sidorov employs a corpus comprising the texts produced by three different authors of the nineteenth century and adopts the support vector machine algorithm to classify n-grams into n-grams of words, part-of-speech tags and characters. The empirical results show the validity of the sn-gram-based

https://doi.org/10.1017/S1351324923000074 Published online by Cambridge University Press

approach in authorship attribution. Chapters 14 and 15 present the ideas of filtered and generalised n-grams, respectively, both essentially looking at the ways of constructing non-continuous sn-grams with different approaches.

In summary, the book exhibits a variety of academic merits, addressing a number of important issues associated with quantitative linguistics and computational linguistics, and showing novel observations and innovative approaches in dealing with n-grams in natural language processing. The fact that this compact volume is only 92 pages long does not detract from its academic value. First, Sidorov notes the limitations in traditional n-gram models regarding the loss of syntactic information and proposes the use of syntactic n-grams in a non-linear structure. Using a wide variety of examples, he shows readers how to comprehend the nature of sn-grams correctly and perform complicated tasks in computational linguistics. Extensive explanations of the rationale pertinent to the construction of sn-gram models pave the way for more in-depth discussions central to dependency syntax, such as sentiment analysis.

Second, Sidorov integrates data processing and computer programming techniques into specific tasks (e.g., authorship attribution) with a critical reflective approach related to classification and clustering. In addition, the source codes for segmenting texts and retrieving n-grams from texts are largely replicable and can be used creatively for future relevant studies. Novice language researchers will be particularly inspired by the novelty of such aspects, moving towards data-intensive scientific discovery.

Third, given that the book is the updated volume of its previous Spanish edition, it includes more examples and case studies in relation to the English language. Both computational proficiency and computational accuracy are taken into account while constructing and training sn-gram models in a dynamic and heterogeneous data environment. In this sense, Sidorov's present work would appeal particularly to those whose expertise is to carry out contrastive analysis between 'genetically distinct' languages.

Overall, in terms of readability, the book features concise academic language and might require readers to have sound basic knowledge about model construction and computer programming skills, although Sidorov notes that no profound knowledge of computing or mathematics is needed. In terms of what this book can offer natural language engineers, as well as quantitative and computational linguistics researchers, it appears that the inclusion of a number of theoretical and methodological innovations fills an important gap in the literature. Such a holistic and thought-provoking monograph would be greatly helpful for providing new algorithms, building new models and exploring new applications. Therefore, the book offers a worthy contribution to the research fields of quantitative linguistics and computational linguistics.

Acknowledgements. This work was supported by National Office for Philosophy and Social Sciences under grant (21BYY046), Fundamental Research Funds for the Central Universities under grant (3132022611) and Liaoning Office for Education Sciences Planning under grant (JG21DB059).

Feng Shi
School of Foreign Languages,
University of Science and Technology Liaoning,
Anshan, P. R. China

Guohua Feng
School of Foreign Languages,
Liaoning Petrochemical College,
Jinzhou, P. R. China
E-mail: 178862677@qq.com

Reference

Szalay A. and Gray J. (2006). Science in an exponential world. *Nature* **440**(7083), 413–414.

https://doi.org/10.1017/S1351324923000074 Published online by Cambridge University Press