SpringerBriefs in Computer Science

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.
Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8-12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

More information about this series at http://www.springer.com/series/10028

Grigori Sidorov

Syntactic n-grams in Computational Linguistics

Springer

Grigori Sidorov
Instituto Politécnico Nacional
Centro de Investigación en Computación
Mexico City, Mexico

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-030-14770-9 ISBN 978-3-030-14771-6 (eBook)
https://doi.org/10.1007/978-3-030-14771-6

Library of Congress Control Number: 2019935141

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This is a new substantially revised edition of the book, where we discuss the use of syntactic information (represented as syntactic n-grams) in the tasks of computational linguistics related to application of machine learning methods. We substantially revised and completed the book adding several new chapters.

The previous edition of this book was published several years ago in Spanish under the name of *Non-linear Construction of N-Grams in Computational Linguistics: Syntactic, Filtered, and Generalized N-Grams*, and we got very positive feedback from many researchers and students, who complimented us saying the book presents really novel idea, describes this idea in a very clear way, and besides presents the background in the very transparent manner, which makes it possible for many persons clarifying the procedures applied in the modern computational linguistics.

We made substantial changes in the book as compared to the previous edition adding several chapters, and we also added more examples in English. We also clarified the relationship between continuous and noncontinuous syntactic n-grams. Let us remind that the concept suggested in the book, syntactic n-grams, is a universal concept, which can be applied in any language.

We are delighted to present to the reader the modified book now in its English version.

This book is about a new approach in the field of computational linguistics related to the idea of constructing n-grams in a nonlinear manner, while the traditional approach consists in using the data from the surface structure of texts, i.e., the linear structure.

In this book, we propose and systematize the concept of syntactic n-grams, which allows using syntactic information within the automatic text processing methods related to classification or clustering. It is a very interesting example of application of linguistic information in the automatic (computational) methods. Roughly speaking, the suggestion is to follow syntactic trees and construct n-grams based on paths in these trees. There are several types of nonlinear n-grams; future work should determine which types of n-grams are more useful in which natural language processing (NLP) tasks. For clarification of the basic concept for the

reader, we dedicate the first part of the book to explanation of basic concepts of computational linguistics and machine learning (vector space model, tf-idf, etc.) and explain the general scheme of design of the experiment in this field.

The book, first and foremost, is intended for specialists in the field of computational linguistics. However, we made an effort to explain in a clear manner how to use n-grams; we provide a large number of examples, and, therefore, we believe that the book is also useful for graduate students who already have some previous background in the field.

We want to emphasize that no profound knowledge of computing or mathematics is required; the proposed concepts are intuitively very clear; we use very few formulas, and if they appear, they are explained in detail.

Mexico City, Mexico Grigori Sidorov

Introduction

In this book, we discuss a novel idea in the field of computational linguistics: the construction of n-grams in a nonlinear manner.

First, we discuss the concept of the vector space model in detail – a conceptual framework for comparison of any type of objects and then its application to the text processing-related tasks, i.e., its use in computational linguistics. Concepts related to word frequency (*tf-idf*) are discussed, and the latent semantic analysis that allows reducing the number of dimensions is briefly presented.

We mention important concepts concerning the design of experiments in computational linguistics and describe the typical scheme of experiment in this area.

We present the concept of traditional (linear) n-grams and compare it with the concept of n-grams obtained in a nonlinear manner: syntactic, filtered, and generalized n-grams.

Syntactic n-grams are n-grams constructed by following paths in syntactic trees. The great advantage of syntactic n-grams is that they allow introducing pure linguistic (syntactic) information into machine learning methods. The disadvantage is that syntactic parsing is required for their construction (note that already there are syntactic parsers available for many languages).

We consider both continuous and noncontinuous syntactic n-grams. When constructing continuous syntactic n-grams, bifurcations (returns, interruptions, ramification) in the syntactic paths are not allowed; when removing this constraint, noncontinuous syntactic n-grams are obtained: sub-trees of length *n* with bifurcations of a syntax tree are considered. It is noteworthy that we can unite these two types of sn-grams: continuous and noncontinuous syntactic n-grams are complete syntactic n-grams; it is all sub-trees of length *n.*

We propose a metalanguage for the representation of noncontinuous syntactic n-grams, i.e., a formal way to represent a noncontinuous syntactic n-gram using brackets and commas, e.g., “a b [c [d, e], f].” In this case, brackets and commas are a part of the sn-grams.

In this book, we also present several examples of construction of continuous and noncontinuous syntactic n-grams for syntactic trees obtained using the FreeLing and the Stanford parsers.

We show that the application of syntactic n-grams in one of the traditional computational linguistics tasks, the task of authorship attribution, gives better results than using traditional n-grams.

Finally, we present several ideas concerning the other types of nonlinearly constructed n-grams:

1. Filtered n-grams: a filter of words or features is built using a certain criterion before constructing n-grams; then n-grams are constructed using only the elements that passed through the filter.
2. Generalized n-grams: words "are generalized" using lexical relations, especially synonymy and hypernymy; in this way, the set of elements used for constructing n-grams is reduced.

Many experimental studies are required in order to determine which construction parameters of continuous and noncontinuous, filtered, and generalized n-grams are the best and for which existing tasks in computational linguistics.

The book systematizes the recent author's proposals on the nonlinear construction of n-grams and their use in vector space model; thereby, some parts of the book are based on the author's previous works published in various journals and conferences with updates and necessary adjustments.

Work is done under the partial support of the Mexican government (CONACYT, SNI); Mexico City government (ICYT-DF PICCO10-120 project); National Polytechnic Institute, Mexico (projects SIP, COFAA); CONACYT project 240844; and FP7 PEOPLE-2010 IRSES: Web Information Quality Evaluation Initiative (WIQ-EI) European Commission project 269180.

Contents