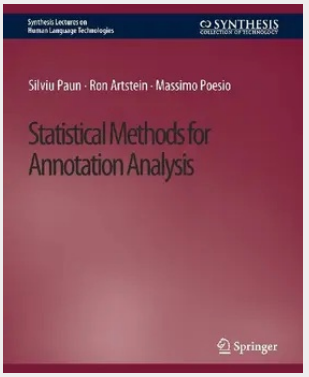
Paun / Artstein / Poesio

Statistical Methods for Annotation Analysis



Labelling data is one of the most fundamental activities in science, and has underpinned practice, particularly in medicine, for decades, as well as research in corpus linguistics since at least the development of the Brown corpus. With the shift towards Machine Learning in Artificial Intelligence (AI), the creation of datasets to be used for training and evaluating AI systems, also known in AI as corpora, has become a central activity in the field as well. Early AI datasets were created on an ad-hoc basis to tackle specific problems. As larger and more reusable datasets were created, requiring greater investment, the need for a more systematic approach to dataset creation arose to ensure increased quality. A range of statistical methods were adopted, often but not exclusively from the medical sciences, to ensure that the labels used were not subjective, or to choose among different labels provided by the coders. A wide variety of such methods is now in regular use. This book is meantto provide a survey of the most widely used among these statistical methods supporting annotation practice. As far as the authors know, this is the first book attempting to cover the two families of methods in wider use. The first family of methods is concerned with the development of labelling schemes and, in particular, ensuring that such schemes are such that sufficient agreement can be observed among the coders. The second family includes methods developed to analyze the output of coders once the scheme has been agreed upon, particularly although not exclusively to identify the most likely label for an item among those provided by the coders. The focus of this book is primarily on Natural Language Processing, the area of AI devoted to the development of models of language interpretation and production, but many if not most of the methods discussed here are also applicable to other areas of AI, or indeed, to other areas of Data Science.

Labelling data is one of the most fundamental activities in science, and has underpinned practice, particularly in medicine, for decades, as well as research in corpus linguistics since at least the development of the Brown corpus. With the shift towards Machine Learning in Artificial Intelligence (AI), the creation of datasets to be used for training and evaluating AI systems, also known in AI as corpora, has become a central activity in the field as well. Early AI datasets were created on an ad-hoc basis to tackle specific problems. As larger and more reusable datasets were created, requiring greater investment, the need for a more systematic approach to dataset creation arose to ensure increased quality. A range of statistical methods were adopted, often but not exclusively from the medical sciences, to ensure that the labels used were not subjective, or to choose among different labels provided by the coders. A wide variety of such methods is now in regular use. This book is meantto provide a survey of the most widely used among these statistical methods supporting annotation practice. As far as the authors know, this is the first book attempting to cover the two families of methods in wider use. The first family of methods is concerned with the development of labelling schemes and, in particular, ensuring that such schemes are such that sufficient agreement can be observed among the coders. The second family includes methods developed to analyze the output of coders once the scheme has been agreed upon, particularly although not exclusively to identify the most likely label for an item among those provided by the coders. The focus of this book is primarily on Natural Language Processing, the area of AI devoted to the development of models of language interpretation and production, but many if not most of the methods discussed here are also applicable to other areas of AI, or indeed, to other areas of Data Science.

69,54 €
64,99 € (zzgl. MwSt.)

Lieferfrist: bis zu 10 Tage

Artikelnummer: 9783031037535
Medium: Buch
ISBN: 978-3-031-03753-5
Verlag: Springer
Erscheinungstermin: 13.01.2022
Sprache(n): Englisch
Auflage: Erscheinungsjahr 2022
Serie: Synthesis Lectures on Human Language Technologies
Produktform: Kartoniert
Gewicht: 420 g
Seiten: 197
Format (B x H): 191 x 235 mm

Kundenservice Verlag Dr. Otto Schmidt KG
Neumannstraße 10, 40235 Düsseldorf | kundenservice@otto-schmidt.de | 0221-93738-997 (Inland)