


A Synoptic Primer: Review of *Wissensrohstoff Text: Eine Einführung in das Text Mining* (2022)

Michael Richter <richter_at_informatik_dot_uni-leipzig_dot_de>, University of Leipzig  <https://orcid.org/0000-0001-7460-4139>

Abstract

Biemann, Heyer, and Quasthoff's *Wissensrohstoff Text* is concerned with conveying a basic understanding of the models and techniques of text mining, as well as insight into which method is procedurally suitable for whichever problem in this area. The book imparts indispensable knowledge, and the new edition makes it possible to describe and discuss the most recent developments in text mining. The book deals also with linguistic fundamentals and with principles of human language processing which is very noteworthy and a unique asset. The book shows in an exemplary way how complex knowledge can be conveyed by means of didactic reduction.

This book is the second and substantially revised edition of the work *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*, published in 2006. With this new edition, the authors seek to take into account the enormous developments within text mining since then. Its target audience is broad: aimed equally at experts and students of computer science, business informatics, digital humanities and linguistics with a focus on or interest in text mining methods. The book contains seven chapters, each preceded by a summary, and the last of which is devoted to sample applications. A great strength of this book (to be further detailed later) is the authors' concern to make their presentation as vivid and lively as possible, using "real world" examples, which immediately make the sense and purpose of any one procedure described clear. Special practical tips and further examples are given in short text blocks after the respective sections under a short "Summary" highlighted in grey. Unfamiliar technical terms can be looked up in a glossary. 1

Chapter 1 introduces the subject matter, that is to say it describes the potential of "text mining" and important concepts and tools. Basic terms and concepts are clarified, such as "information", i.e. data in a context, and "knowledge" in the sense of procedural and action-oriented knowledge for problem solving. Much space is given to the resource, the text. The reader learns the structure-principle of texts, the distinction of types and tokens, and the phenomena of redundancy in texts on an information structure level and of congruence on a linguistic level. The chapter then discusses and justifies the basic question of why linguistic knowledge in the form of language models is necessary for automatic natural language processing and thus prepares chapter 2. The last section of chapter 1 introduces linguistic structures of texts from letters to sentences, levels of representation and processing of language, and the two main paradigms of rule-based and statistical approaches of describing linguistic regularities, thereby preparing the ground for further elaboration in the later chapters. That is followed by an introduction to statistical approaches to text analysis, with – first – a reference to the types of unsupervised learning. The text then differentiates between frequentist and Bayesian approaches, the latter being used in machine learning procedures in text mining as well as in the specific application area of topic models. Finally, the paradigm shift from rule- and pattern-based approaches to statistical methods is discussed in detail. Whereas in the 1960s it was procedures based on automata theory, since the 1990s data-driven statistical procedures have dominated in automatic language processing, which are also the basis of modern neural network language processing. 2

Chapter 2 deals with linguistic foundations, and the discussion starts with a structuralist theoretical framework. The 3

linguistic levels of morphology, syntax and semantics are described, and the highly influential formalism of phrase structure grammars is presented that Noam Chomsky introduced into linguistics, drawing inspiration from automata theory. As an alternative rule-based syntax model, Lucien Tesnière's Dependency Grammar is discussed, that, in contrast to Chomsky's formalism, interprets the verb as the central element of the sentence. Phrase structure rules can be enriched by probabilities, as the text exemplifies with a grammar fragment. Such work assigns probabilities to parse trees.

The sections on semantics give an overview of basic paradigms of semantic modelling, each of which is presented in detail but concisely and limited to the essentials, that is: procedural, referential semantics and structuralistic distributional semantics. A semantics of the first type interprets meaning in terms of the processes and actions triggered by linguistic stimuli. Referential semantics is usually a use case of predicate logic and formalises the relationship between a linguistic expression and the "world" represented as a model. Currently, the most influential and widely used model of semantics is based on the distributional hypothesis: that from the context of a word its meaning can be inferred, or in other words, words with similar contexts must have similar meanings. Hereby, the text brings state-of-the-art methods into focus, such as word embeddings, a subject that is then taken up again in chapter 5. Special attention is given to the semantic relations between words such as transitive relations and hierarchical relationships which mirror the hierarchical structure of ontologies in the mental lexicon that is commonly represented as (directed or undirected) graphs. The chapter is completed with explanations on technical texts and on the relevance of specific terminology, as well as on how to deal with exceptions in language.

Chapter 3 addresses the automatic processing of text, i.e., the operational implementation through models of different characters. A distinction is made between rule-based, statistical and neural methods. The section on neural networks spans an arc from the different types of neural networks with their respective characterising architectures to modern state-of-the-art models and procedures such as *contextualised word embeddings* and *transformer models*, which are becoming increasingly popular for instance in automatic translation and text generation. The chapter then deals in detail with processing pipelines, i.e. the "chaining" of processes for text processing, starting with segmentation at word and sentence levels to morphologic, syntactic and, finally, semantic processing. Essential applications such as named entity recognition, sentiment analysis, open information extraction, terminology extraction, and entity-centred retrieval are discussed and accompanied by illustrative examples from practice. The chapter concludes by addressing the important and practical issue of scaling, which is defined as the increase in performance of technical infrastructures in the face of growing requirements, i.e. constantly increasing amounts of available data, and parallelisation is discussed as a commonly used scaling-measure.

Text mining is based on linguistic corpora, which is the subject of chapter 4. Essentially, different types of corpora are presented, taking as points of departure the dichotomy of generic, i.e. already existing corpora, and self-created, i.e. customised, tailor-made corpora for one's own research question. Subsequently, the text deals with the problems and pitfalls of compiling one's own corpus, such as bias in any form. An option available to today's researchers is the creation of corpora based on web "crawling." The requirement of quality assurance is addressed in detail, for example, how to identify and remove erroneous sentences, sentences in a language other than the corpus language, and how to remove duplicates and quasi-duplicates. As befits the action-oriented nature of the book, the reader is also provided with a number of concrete rules for "cleaning" corpora and ranking sentences that have proven effective. The chapter then provides an overview of formats for storing corpora such as XML, JSON, and the CoNLL / CoNLL-U column format of Universal Dependencies corpora. The subsequent sections are devoted to the question of how to query the corpora as efficiently as possible, and the chapter concludes with the treatment of purely lexical resources, i.e. lists of words, represented a.o. (among others) as graphs such as in WordNet.

Chapter 5 deals with the elementary field of language statistics, i.e. with methods and techniques for detecting statistical frequencies such as type-token ratio or mean word length in texts and evaluating linguistic phenomena and productions on that basis. A prominent place in this chapter is occupied by Zipf's laws and principles that model processes of speech production and comprehension in terms of information theory, linguistic processes that are subject to the behavioural economy principle of least effort. In addition, the Zipfian laws allow languages and linguistic productions to be characterised by key indicators such as constants, and allow the production and the processing of language, to be

4

5

6

7

modelled as a process of human behaviour.

The chapter then discusses in detail the distributional model, which is based on the co-occurrences of words to be modelled linguistically. That model is a cornerstone of computational semantics, and its basic ideas are also relevant for state-of-the-art network models such as (contextualised) word embeddings. In the case of co-occurrences, significance measures are introduced that can be used to decide whether particular patterns occur more frequently than can be statistically expected. An outline of distributional semantics and of probabilistic language models follows: word embeddings represent words and texts by vectors with a high number of entries and play a crucial role in the language processing and generation by neural networks. The section on word embeddings is concluded by describing contextualised word embeddings in transformer models such as BERT, which use attention and feed-forward layers to predict multiple levels of context representation. The subsequent sections illustrate how linguistic statistics can be utilised to disambiguate word meanings and to check the quality of corpora. The chapter concludes with a description of some in-depth work on the comparison of texts and corpora by means of difference analyses, i.e., for example, by comparisons of distributions and contexts of words.

8

Chapter 6 deals with machine learning, a central process in text mining. In order for machines to learn, to evaluate, and to classify human language and its productions, features of linguistic productions must be extracted and represented in a machine-processable way. The text first introduces procedures of unsupervised learning, i.e. the classification procedures of clustering and their characterising features. The representation of words, sentences, texts, etc. as vectors is immediately apparent, since mathematical distance measures enable the exact determination of proximity and distance of words and texts. This is demonstrated by example applications of clustering, such as hierarchical clustering of word types, induction of word meaning, and document classification. The explanations on clustering are rounded off by a section on the evaluation of cluster quality. Topic detection in documents by the probabilistic generative method of Latent Dirichlet Allocation is an unsupervised learning method as well. The algorithm generates texts as a distribution of topics, and topics as a distribution of words present in the text. In an impressively didactic reduced manner and limited to the essential principles of the procedure, the reader is carefully introduced to this challenging topic.

9

The chapter then turns to supervised learning procedures in classifications and discusses first the general procedure of supervised classification, including splitting of the data set into training and test sets, the tuning of the classification algorithm for instance by means of hyperparameters, and the avoidance of the danger of overfitting, for example in classification by neural networks by the early-stop method. Before turning to classification methods using neural networks, the text describes the classical Naïve Bayes approach and decision trees. Subsequently, the relevance of the annotation of training data in supervised learning is taken into account, followed by a brief and illustrative treatment of hidden Markov models and the Viterbi algorithm for the classification of sequences. The chapter too is rounded off by a presentation of the evaluation possibilities of classifications, on the one hand through quantity-based measures such as accuracy, precision, recall, and F1, and on the other hand through neural evaluation methods such as end-to-end learning and transfer learning.

10

Finally, Chapter 7 is a concluding illustration of the body of knowledge through a detailed description of studies, some from the most current research context such as a study on context volatility. Furthermore, the chapter describes studies on terminology extraction, proper noun search, sentiment analysis, trend analysis and neologisms.

11

In summary, this work offers a didactically excellent, comprehensive presentation of the state-of-the-art in text mining and information retrieval. The book is equally applicable as seminar and lecture material as well as for self-study. The target groups are novices as well as professionals in the field of automatic language processing.

12

Works Cited

Biemann, Heyer, and Quasthoff 2022 Biemann, C., Heyer, G. and Quasthoff, U. (2022) *Wissensrohstoff Text: Eine Einführung in das Text Mining*. Revised Second Edition. Springer Vieweg Wiesbaden.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.