Introduction To Linguistic Annotation And Text Analytics Graham Wilcock

Yeah, reviewing a book Introduction To Linguistic Annotation And Text Analytics Graham Wilcock could grow your near links listings. This is just one of the solutions for you to be successful. As understood, endowment does not recommend that you have astounding points.

Comprehending as well as covenant even more than new will come up with the money for each success. adjacent to, the notice as without difficulty as acuteness of this Introduction To Linguistic Annotation And Text Analytics Graham Wilcock can be taken as skillfully as picked to act.

Aspects of Automatic Text Analysis Alexander Mehler 2007 This book presents recent developments in automatic text analysis. Providing an overview of linguistic modeling, it collects contributions of authors from a multidisciplinary area that focus on the topic of automatic text analysis from different perspectives. It includes chapters on cognitive modeling and visual systems modeling, and contributes to the computational linguistic and information theoretical grounding of automatic text analysis.
Statistical Significance Testing for Natural Language Processing Rotem Dror 2022-06-01 Data-driven experimental analysis has become the main evaluation tool of Natural Language Processing (NLP) algorithms. In fact, in the last decade, it has become rare to see an NLP paper, particularly one that proposes a new algorithm, that does not include extensive experimental analysis, and the number of involved tasks, datasets, domains, and languages is constantly growing. This emphasis on empirical results highlights the role of statistical significance testing in NLP research: If we, as a community, rely on empirical evaluation to validate our hypotheses and reveal the correct language processing mechanisms, we better be sure that our results are not coincidental. The goal of this book is to discuss the main aspects of statistical significance testing in NLP. Our guiding assumption throughout the book is that the basic question NLP researchers and engineers deal with is whether or not one algorithm can be considered better than another one. This question drives the field forward as it allows the constant progress of developing better technology for language processing challenges. In practice, researchers and engineers would like to draw the right conclusion from a limited set of experiments, and this conclusion should hold for other experiments with datasets they do not have at their disposal or that they cannot perform due to limited time and resources. The book hence discusses the opportunities and challenges in using statistical significance testing in NLP, from the point of view of experimental comparison between two algorithms. We cover topics such as choosing an appropriate significance test for the major NLP tasks, dealing with the unique aspects of significance testing for non-convex deep neural networks, accounting for a large number of comparisons between two NLP algorithms in a statistically valid manner (multiple hypothesis testing), and, finally, the unique challenges yielded by the nature of the data and practices of the field.
The Oxford Handbook of Computational Linguistics Ruslan Mitkov 2022-06-02 Ruslan Mitkov's highly successful Oxford Handbook of Computational Linguistics has been substantially revised and expanded in this second edition. Alongside updated accounts of the topics covered in the first edition, it includes 17 new chapters on subjects such as semantic role-labelling, text-to-speech synthesis, translation technology, opinion mining and sentiment analysis, and the application of Natural Language Processing in educational and biomedical contexts, among many others. The volume is divided into four parts that examine, respectively: the linguistic fundamentals of computational linguistics; the methods and resources used, such as statistical modelling, machine learning, and corpus annotation; key language processing tasks including text segmentation, anaphora resolution, and speech recognition; and the major applications of Natural Language Processing, from machine translation to author profiling. The book will be an essential reference for researchers and students in computational linguistics and Natural Language Processing, as well as those working in related industries.
Validity, Reliability, and Significance Stefan Riezler 2022-06-01 Empirical methods are means to answering methodological questions of empirical sciences by statistical techniques. The methodological questions addressed in this book include the problems of validity, reliability, and significance. In the case of machine learning, these correspond to the questions of whether a model predicts what it purports to predict, whether a model's performance is consistent across replications, and whether a performance difference between two models is due to chance, respectively. The goal of this book is to answer these questions by concrete statistical tests that can be applied to assess validity, reliability, and significance of data annotation and machine learning prediction in the fields of NLP and data science. Our focus is on model-based empirical methods where data annotations and model predictions are treated as training data for interpretable probabilistic models from the well-understood families of generalized additive models (GAMs) and linear mixed effects models (LMEMs). Based on the interpretable parameters of the trained GAMs or LMEMs, the book presents model-based statistical tests such as a validity test that allows detecting circular features that circumvent learning. Furthermore, the book discusses a reliability coefficient using variance decomposition based on random effect parameters of LMEMs. Last, a significance test based on the likelihood ratio of nested LMEMs trained on the performance scores of two machine learning models is shown to naturally allow the inclusion of variations in meta-parameter settings into hypothesis testing, and further facilitates a refined system comparison conditional on properties of input data. This book can be used as an introduction to empirical methods for machine learning in general, with a special focus on applications in NLP and data science. The book is self-contained, with an appendix on the mathematical background on GAMs and LMEMs, and with an accompanying webpage including R code to replicate experiments presented in the book.
Linguistic Fundamentals for Natural Language Processing Emily M. Bender 2022-05-31 Many NLP tasks have at their core a subtask of extracting the dependencies—who did what to whom—from natural language sentences. This task can be understood as the inverse of the problem solved in different ways by diverse human languages, namely, how to indicate the relationship between different parts of a sentence. Understanding how languages solve the problem can be extremely useful in both feature design and error analysis in the application of machine learning to NLP. Likewise, understanding cross-linguistic variation can be important for the design of MT systems and other multilingual applications. The purpose of this book is to present in a succinct and accessible fashion information about the morphological and syntactic structure of human languages that can be useful in creating more linguistically sophisticated, more language-independent, and thus more successful NLP systems. Table of Contents: Acknowledgments / Introduction/motivation / Morphology: Introduction / Morphophonology / Morphosyntax / Syntax: Introduction / Parts of speech / Heads, arguments, and adjuncts /

Argument types and grammatical functions / Mismatches between syntactic position and semantic roles / Resources / Bibliography / Author's Biography / General Index / Index of Languages

Embeddings in Natural Language Processing Mohammad Taher Pilehvar 2022-05-31 Embeddings have undoubtedly been one of the most influential research areas in Natural Language Processing (NLP). Encoding information into a low-dimensional vector representation, which is easily integrable in modern machine learning models, has played a central role in the development of NLP. Embedding techniques initially focused on words, but the attention soon started to shift to other forms: from graph structures, such as knowledge bases, to other types of textual content, such as sentences and documents. This book provides a high-level synthesis of the main embedding techniques in NLP, in the broad sense. The book starts by explaining conventional word vector space models and word embeddings (e.g., Word2Vec and GloVe) and then moves to other types of embeddings, such as word sense, sentence and document, and graph embeddings. The book also provides an overview of recent developments in contextualized representations (e.g., ELMo and BERT) and explains their potential in NLP. Throughout the book, the reader can find both essential information for understanding a certain topic from scratch and a broad overview of the most successful techniques developed in the literature.

Corpus Annotation R. G. Garside 2016-07-27 Corpus Annotation gives an up-to-date picture of this fascinating new area of research, and will provide essential reading for newcomers to the field as well as those already involved in corpus annotation. Early chapters introduce the different levels and techniques of corpus annotation. Later chapters deal with software developments, applications, and the development of standards for the evaluation of corpus annotation. While the book takes detailed account of research world-wide, its focus is particularly on the work of the UCREL (University Centre for Computer Corpus Research on Language) team at Lancaster University, which has been at the forefront of developments in the field of corpus annotation since its beginnings in the 1970s.

Sentiment Analysis and Opinion Mining Bing Liu 2022-05-31 Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations. This book is a comprehensive introductory and survey text. It covers all important topics and the latest developments in the field with over 400 references. It is suitable for students, researchers and practitioners who are interested in social media analysis in general and sentiment analysis in particular. Lecturers can readily use it in class for courses on natural language processing, social media analysis, text mining, and data mining. Lecture slides are also available online. Table of Contents: Preface / Sentiment Analysis: A Fascinating Problem / The Problem of Sentiment Analysis / Document Sentiment Classification / Sentence Subjectivity and Sentiment Classification / Aspect-Based Sentiment Analysis / Sentiment Lexicon Generation / Opinion Summarization / Analysis of Comparative Opinions / Opinion Search and Retrieval / Opinion Spam Detection / Quality of Reviews / Concluding Remarks / Bibliography / Author Biography

Data-Intensive Text Processing with MapReduce Jimmy Lin 2022-05-31 Our world is being revolutionized by data-driven methods: access to large amounts of data has generated new insights and opened exciting new opportunities in commerce, science, and computing applications. Processing the enormous quantities of data necessary for these advances requires large clusters, making distributed computing paradigms more crucial than ever. MapReduce is a programming model for expressing distributed computations on massive datasets and an execution framework for large-scale data processing on clusters of commodity servers. The programming model provides an easy-to-understand abstraction for designing scalable algorithms, while the execution framework transparently handles many system-level details, ranging from scheduling to synchronization to fault tolerance. This book focuses on MapReduce algorithm design, with an emphasis on text processing algorithms common in natural language processing, information retrieval, and machine learning. We introduce the notion of MapReduce design patterns, which represent general reusable solutions to commonly occurring problems across a variety of problem domains. This book not only intends to help the reader "think in MapReduce", but also discusses limitations of the programming model as well. Table of Contents: Introduction / MapReduce Basics / MapReduce Algorithm Design / Inverted Indexing for Text Retrieval / Graph Algorithms / EM Algorithms for Text Processing / Closing Remarks

Cross-Lingual Word Embeddings Anders Søgaard 2022-05-31 The majority of natural language processing (NLP) is English language processing, and while there is good language technology support for (standard varieties of) English, support for Albanian, Burmese, or Cebuano--and most other languages--remains limited. Being able to bridge this digital divide is important for scientific and democratic reasons but also represents an enormous growth potential. A key challenge for this to happen is learning to align basic meaning-bearing units of different languages. In this book, the authors survey and discuss recent and historical work on supervised and unsupervised learning of such alignments. Specifically, the book focuses on so-called cross-lingual word embeddings. The survey is intended to be systematic, using consistent notation and putting the available methods on comparable form, making it easy to compare wildly different approaches. In so doing, the authors establish previously unreported relations between these methods and are able to present a fast-growing literature in a very compact way. Furthermore, the authors discuss how best to evaluate cross-lingual word embedding methods and survey the resources available for students and researchers interested in this topic.

Bayesian Analysis in Natural Language Processing, Second Edition Shay Cohen 2022-05-31 Natural language processing (NLP) went through a profound transformation in the mid-1980s when it shifted to make heavy use of corpora and data-driven techniques to analyze language. Since then, the use of statistical techniques in NLP has evolved in several ways. One such example of evolution took place in the late 1990s or early 2000s, when full-fledged Bayesian machinery was introduced to NLP. This Bayesian approach to NLP has come to accommodate various shortcomings in the frequentist approach and to enrich it, especially in the unsupervised setting, where statistical learning is done without target prediction examples. In this book, we cover the methods and algorithms that are needed to fluently read Bayesian learning papers in NLP and to do research in the area. These methods and algorithms are partially borrowed from both machine learning and statistics and are partially developed "in-house" in NLP. We cover inference techniques such as Markov chain Monte Carlo sampling and variational inference, Bayesian estimation, and nonparametric modeling. In response to rapid changes in the field, this second edition of the book includes a new chapter on representation learning and neural networks in the Bayesian context. We also cover fundamental concepts in Bayesian statistics such as prior distributions, conjugacy, and generative modeling. Finally, we review some of the fundamental modeling techniques in NLP, such as grammar modeling, neural networks and representation learning, and their use with Bayesian analysis.

The Semantic Web: Research and Applications Elena Simperl 2012-05-24 This book constitutes the refereed proceedings of the 9th

Extended Semantic Web Conference, ESWC 2012, held in Heraklion, Crete, Greece, in May 2012. The 53 revised full papers presented were carefully reviewed and selected from 212 submissions. They are organized in tracks on linked open data, machine learning, natural language processing and information retrieval, ontologies, reasoning, semantic data management, services, processes, and cloud computing, social Web and Web science, in-use and industrial, digital libraries and cultural heritage, and e-government. The book also includes 13 PhD papers presented at the PhD Symposium.

Semantic Similarity from Natural Language and Ontology Analysis Sébastien Harispe 2022-05-31 Artificial Intelligence federates numerous scientific fields in the aim of developing machines able to assist human operators performing complex treatments---most of which demand high cognitive skills (e.g. learning or decision processes). Central to this quest is to give machines the ability to estimate the likeness or similarity between things in the way human beings estimate the similarity between stimuli. In this context, this book focuses on semantic measures: approaches designed for comparing semantic entities such as units of language, e.g. words, sentences, or concepts and instances defined into knowledge bases. The aim of these measures is to assess the similarity or relatedness of such semantic entities by taking into account their semantics, i.e. their meaning---intuitively, the words tea and coffee, which both refer to stimulating beverage, will be estimated to be more semantically similar than the words toffee (confection) and coffee, despite that the last pair has a higher syntactic similarity. The two state-of-the-art approaches for estimating and quantifying semantic similarities/relatedness of semantic entities are presented in detail: the first one relies on corpora analysis and is based on Natural Language Processing techniques and semantic models while the second is based on more or less formal, computer-readable and workable forms of knowledge such as semantic networks, thesauri or ontologies. Semantic measures are widely used today to compare units of language, concepts, instances or even resources indexed by them (e.g., documents, genes). They are central elements of a large variety of Natural Language Processing applications and knowledge-based treatments, and have therefore naturally been subject to intensive and interdisciplinary research efforts during last decades. Beyond a simple inventory and categorization of existing measures, the aim of this monograph is to convey novices as well as researchers of these domains toward a better understanding of semantic similarity estimation and more generally semantic measures. To this end, we propose an in-depth characterization of existing proposals by discussing their features, the assumptions on which they are based and empirical results regarding their performance in particular applications. By answering these questions and by providing a detailed discussion on the foundations of semantic measures, our aim is to give the reader key knowledge required to: (i) select the more relevant methods according to a particular usage context, (ii) understand the challenges offered to this field of study, (iii) distinguish room of improvements for state-of-the-art approaches and (iv) stimulate creativity toward the development of new approaches. In this aim, several definitions, theoretical and practical details, as well as concrete applications are presented.

Computational Methods for Corpus Annotation and Analysis Xiaofei Lu 2014-07-08 In the past few decades the use of increasingly large text corpora has grown rapidly in language and linguistics research. This was enabled by remarkable strides in natural language processing (NLP) technology, technology that enables computers to automatically and efficiently process, annotate and analyze large amounts of spoken and written text in linguistically and/or pragmatically meaningful ways. It has become more desirable than ever before for language and linguistics researchers who use corpora in their research to gain an adequate understanding of the relevant NLP technology to take full advantage of its capabilities. This volume provides language and linguistics researchers with an accessible introduction to the state-of-the-art NLP technology that facilitates automatic annotation and analysis of large text corpora at both shallow and deep linguistic levels. The book covers a wide range of computational tools for lexical, syntactic, semantic, pragmatic and discourse analysis, together with detailed instructions on how to obtain, install and use each tool in different operating systems and platforms. The book illustrates how NLP technology has been applied in recent corpus-based language studies and suggests effective ways to better integrate such technology in future corpus linguistics research. This book provides language and linguistics researchers with a valuable reference for corpus annotation and analysis.

Finite-State Text Processing Kyle Gorman 2022-06-01 Weighted finite-state transducers (WFSTs) are commonly used by engineers and computational linguists for processing and generating speech and text. This book first provides a detailed introduction to this formalism. It then introduces Pynini, a Python library for compiling finite-state grammars and for combining, optimizing, applying, and searching finite-state transducers. This book illustrates this library's conventions and use with a series of case studies. These include the compilation and application of context-dependent rewrite rules, the construction of morphological analyzers and generators, and text generation and processing applications.

Bayesian Analysis in Natural Language Processing Shay Cohen 2022-11-10 Natural language processing (NLP) went through a profound transformation in the mid-1980s when it shifted to make heavy use of corpora and data-driven techniques to analyze language. Since then, the use of statistical techniques in NLP has evolved in several ways. One such example of evolution took place in the late 1990s or early 2000s, when full-fledged Bayesian machinery was introduced to NLP. This Bayesian approach to NLP has come to accommodate for various shortcomings in the frequentist approach and to enrich it, especially in the unsupervised setting, where statistical learning is done without target prediction examples. We cover the methods and algorithms that are needed to fluently read Bayesian learning papers in NLP and to do research in the area. These methods and algorithms are partially borrowed from both machine learning and statistics and are partially developed "in-house" in NLP. We cover inference techniques such as Markov chain Monte Carlo sampling and variational inference, Bayesian estimation, and nonparametric modeling. We also cover fundamental concepts in Bayesian statistics such as prior distributions, conjugacy, and generative modeling. Finally, we cover some of the fundamental modeling techniques in NLP, such as grammar modeling and their use with Bayesian analysis.

Statistical Methods for Annotation Analysis Silviu Paun 2022-05-31 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 meant to 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.

Natural Language Processing for Historical Texts Michael Piotrowski 2022-05-31 More and more historical texts are becoming available in digital form. Digitization of paper documents is motivated by the aim of preserving cultural heritage and making it more accessible, both to laypeople and scholars. As digital images cannot be searched for text, digitization projects increasingly strive to create digital text, which can be searched and otherwise automatically processed, in addition to facsimiles. Indeed, the emerging field of digital humanities heavily relies on the availability of digital text for its studies. Together with the increasing availability of historical texts in digital form, there is a growing interest in applying natural language processing (NLP) methods and tools to historical texts. However, the specific linguistic properties of historical texts -- the lack of standardized orthography, in particular -- pose special challenges for NLP. This book aims to give an introduction to NLP for historical texts and an overview of the state of the art in this field. The book starts with an overview of methods for the acquisition of historical texts (scanning and OCR), discusses text encoding and annotation schemes, and presents examples of corpora of historical texts in a variety of languages. The book then discusses specific methods, such as creating part-of-speech taggers for historical languages or handling spelling variation. A final chapter analyzes the relationship between NLP and the digital humanities. Certain recently emerging textual genres, such as SMS, social media, and chat messages, or newsgroup and forum postings share a number of properties with historical texts, for example, nonstandard orthography and grammar, and profuse use of abbreviations. The methods and techniques required for the effective processing of historical texts are thus also of interest for research in other domains. Table of Contents: Introduction / NLP and Digital Humanities / Spelling in Historical Texts / Acquiring Historical Texts / Text Encoding and Annotation Schemes / Handling Spelling Variation / NLP Tools for Historical Languages / Historical Corpora / Conclusion / Bibliography

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition Hang Li 2022-05-31 Learning to rank refers to machine learning techniques for training a model in a ranking task. Learning to rank is useful for many applications in information retrieval, natural language processing, and data mining. Intensive studies have been conducted on its problems recently, and significant progress has been made. This lecture gives an introduction to the area including the fundamental problems, major approaches, theories, applications, and future work. The author begins by showing that various ranking problems in information retrieval and natural language processing can be formalized as two basic ranking tasks, namely ranking creation (or simply ranking) and ranking aggregation. In ranking creation, given a request, one wants to generate a ranking list of offerings based on the features derived from the request and the offerings. In ranking aggregation, given a request, as well as a number of ranking lists of offerings, one wants to generate a new ranking list of the offerings. Ranking creation (or ranking) is the major problem in learning to rank. It is usually formalized as a supervised learning task. The author gives detailed explanations on learning for ranking creation and ranking aggregation, including training and testing, evaluation, feature creation, and major approaches. Many methods have been proposed for ranking creation. The methods can be categorized as the pointwise, pairwise, and listwise approaches according to the loss functions they employ. They can also be categorized according to the techniques they employ, such as the SVM based, Boosting based, and Neural Network based approaches. The author also introduces some popular learning to rank methods in details. These include: PRank, OC SVM, McRank, Ranking SVM, IR SVM, GBRank, RankNet, ListNet & ListMLE, AdaRank, SVM MAP, SoftRank, LambdaRank, LambdaMART, Borda Count, Markov Chain, and CRanking. The author explains several example applications of learning to rank including web search, collaborative filtering, definition search, keyphrase extraction, query dependent summarization, and re-ranking in machine translation. A formulation of learning for ranking creation is given in the statistical learning framework. Ongoing and future research directions for learning to rank are also discussed. Table of Contents: Learning to Rank / Learning for Ranking Creation / Learning for Ranking Aggregation / Methods of Learning to Rank / Applications of Learning to Rank / Theory of Learning to Rank / Ongoing and Future Work

Cross-Language Information Retrieval Jian-Yun Nie 2022-05-31 Search for information is no longer exclusively limited within the native language of the user, but is more and more extended to other languages. This gives rise to the problem of cross-language information retrieval (CLIR), whose goal is to find relevant information written in a different language to a query. In addition to the problems of monolingual information retrieval (IR), translation is the key problem in CLIR: one should translate either the query or the documents from a language to another. However, this translation problem is not identical to full-text machine translation (MT): the goal is not to produce a human-readable translation, but a translation suitable for finding relevant documents. Specific translation methods are thus required. The goal of this book is to provide a comprehensive description of the specific problems arising in CLIR, the solutions proposed in this area, as well as the remaining problems. The book starts with a general description of the monolingual IR and CLIR problems. Different classes of approaches to translation are then presented: approaches using an MT system, dictionary-based translation and approaches based on parallel and comparable corpora. In addition, the typical retrieval effectiveness using different approaches is compared. It will be shown that translation approaches specifically designed for CLIR can rival and outperform high-quality MT systems. Finally, the book offers a look into the future that draws a strong parallel between query expansion in monolingual IR and query translation in CLIR, suggesting that many approaches developed in monolingual IR can be adapted to CLIR. The book can be used as an introduction to CLIR. Advanced readers can also find more technical details and discussions about the remaining research challenges in the future. It is suitable to new researchers who intend to carry out research on CLIR. Table of Contents: Preface / Introduction / Using Manually Constructed Translation Systems and Resources for CLIR / Translation Based on Parallel and Comparable Corpora / Other Methods to Improve CLIR / A Look into the Future: Toward a Unified View of Monolingual IR and CLIR? / References / Author Biography

Semantic Relations Between Nominals Vivi Nastase 2013-04-26 People make sense of a text by identifying the semantic relations which connect the entities or concepts described by that text. A system which aspires to human-like performance must also be equipped to identify, and learn from, semantic relations in the texts it processes. Understanding even a simple sentence such as "Opportunity and Curiosity find similar rocks on Mars" requires recognizing relations (rocks are located on Mars, signalled by the word on) and drawing on already known relations (Opportunity and Curiosity are instances of the class of Mars rovers). A language-understanding system should be able to find such relations in documents and progressively build a knowledge base or even an ontology. Resources of this kind assist continuous learning and other advanced language-processing tasks such as text summarization, question answering and machine translation. The book discusses the recognition in text of semantic relations which capture interactions between base noun phrases. After a brief historical background, we introduce a range of relation inventories of varying granularity, which have been proposed by computational linguists. There is also variation in the scale at which systems operate, from snippets all the way to the whole Web, and in the techniques of recognizing relations in texts, from full supervision through weak or distant supervision to self-supervised or completely unsupervised methods. A discussion of supervised learning covers available datasets, feature sets which describe relation instances, and successful algorithms. An overview of weakly supervised and unsupervised learning zooms in on the acquisition of relations from large corpora with hardly any annotated data. We show how bootstrapping from seed examples or patterns scales up to very large text

collections on the Web. We also present machine learning techniques in which data redundancy and variability lead to fast and reliable relation extraction.

Automated Essay Scoring Beata Beigman Klebanov 2022-05-31 This book discusses the state of the art of automated essay scoring, its challenges and its potential. One of the earliest applications of artificial intelligence to language data (along with machine translation and speech recognition), automated essay scoring has evolved to become both a revenue-generating industry and a vast field of research, with many subfields and connections to other NLP tasks. In this book, we review the developments in this field against the backdrop of Elias Page's seminal 1966 paper titled "The Imminence of Grading Essays by Computer." Part 1 establishes what automated essay scoring is about, why it exists, where the technology stands, and what are some of the main issues. In Part 2, the book presents guided exercises to illustrate how one would go about building and evaluating a simple automated scoring system, while Part 3 offers readers a survey of the literature on different types of scoring models, the aspects of essay quality studied in prior research, and the implementation and evaluation of a scoring engine. Part 4 offers a broader view of the field inclusive of some neighboring areas, and Part \ref{part5} closes with summary and discussion. This book grew out of a week-long course on automated evaluation of language production at the North American Summer School for Logic, Language, and Information (NASSLLI), attended by advanced undergraduates and early-stage graduate students from a variety of disciplines. Teachers of natural language processing, in particular, will find that the book offers a useful foundation for a supplemental module on automated scoring. Professionals and students in linguistics, applied linguistics, educational technology, and other related disciplines will also find the material here useful.

Natural Language Processing for Social Media Atefeh Farzindar 2015-08-31 In recent years, online social networking has revolutionized interpersonal communication. The newer research on language analysis in social media has been increasingly focusing on the latter's impact on our daily lives, both on a personal and a professional level. Natural language processing (NLP) is one of the most promising avenues for social media data processing. It is a scientific challenge to develop powerful methods and algorithms which extract relevant information from a large volume of data coming from multiple sources and languages in various formats or in free form. We discuss the challenges in analyzing social media texts in contrast with traditional documents. Research methods in information extraction, automatic categorization and clustering, automatic summarization and indexing, and statistical machine translation need to be adapted to a new kind of data. This book reviews the current research on Natural Language Processing (NLP) tools and methods for processing the non-traditional information from social media data that is available in large amounts (big data), and shows how innovative NLP approaches can integrate appropriate linguistic information in various fields such as social media monitoring, health care, business intelligence, industry, marketing, and security and defense. We review the existing evaluation metrics for NLP and social media applications, and the new efforts in evaluation campaigns or shared tasks on new datasets collected from social media. Such tasks are organized by the Association for Computational Linguistics (such as SemEval tasks) or by the National Institute of Standards and Technology via the Text REtrieval Conference (TREC) and the Text Analysis Conference (TAC). In the concluding chapter, we discuss the importance of this dynamic discipline and its great potential for NLP in the coming decade, in the context of changes in mobile technology, cloud computing, and social networking.

Argumentation Mining Manfred Stede 2022-06-01 Argumentation mining is an application of natural language processing (NLP) that emerged a few years ago and has recently enjoyed considerable popularity, as demonstrated by a series of international workshops and by a rising number of publications at the major conferences and journals of the field. Its goals are to identify argumentation in text or dialogue; to construct representations of the constellation of claims, supporting and attacking moves (in different levels of detail); and to characterize the patterns of reasoning that appear to license the argumentation. Furthermore, recent work also addresses the difficult tasks of evaluating the persuasiveness and quality of arguments. Some of the linguistic genres that are being studied include legal text, student essays, political discourse and debate, newspaper editorials, scientific writing, and others. The book starts with a discussion of the linguistic perspective, characteristics of argumentative language, and their relationship to certain other notions such as subjectivity. Besides the connection to linguistics, argumentation has for a long time been a topic in Artificial Intelligence, where the focus is on devising adequate representations and reasoning formalisms that capture the properties of argumentative exchange. It is generally very difficult to connect the two realms of reasoning and text analysis, but we are convinced that it should be attempted in the long term, and therefore we also touch upon some fundamentals of reasoning approaches. Then the book turns to its focus, the computational side of mining argumentation in text. We first introduce a number of annotated corpora that have been used in the research. From the NLP perspective, argumentation mining shares subtasks with research fields such as subjectivity and sentiment analysis, semantic relation extraction, and discourse parsing. Therefore, many technical approaches are being borrowed from those (and other) fields. We break argumentation mining into a series of subtasks, starting with the preparatory steps of classifying text as argumentative (or not) and segmenting it into elementary units. Then, central steps are the automatic identification of claims, and finding statements that support or oppose the claim. For certain applications, it is also of interest to compute a full structure of an argumentative constellation of statements. Next, we discuss a few steps that try to 'dig deeper': to infer the underlying reasoning pattern for a textual argument, to reconstruct unstated premises (so-called 'enthymemes'), and to evaluate the quality of the argumentation. We also take a brief look at 'the other side' of mining, i.e., the generation or synthesis of argumentative text. The book finishes with a summary of the argumentation mining tasks, a sketch of potential applications, and a--necessarily subjective--outlook for the field.

Introduction to Linguistic Annotation and Text Analytics Graham Wilcock 2009 formats using XSLT transformations. The two main text analytics architectures, GATE and UIMA, are then described and compared, with practical exercises showing how to configure and customize them. The final chapter is an introduction to text analytics, describing the main applications and functions including named entity recognition, coreference resolution and information extraction, with practical examples using both open source and commercial tools." --Book Jacket.

Discourse Processing Manfred Stede 2022-06-01 Discourse Processing here is framed as marking up a text with structural descriptions on several levels, which can serve to support many language-processing or text-mining tasks. We first explore some ways of assigning structure on the document level: the logical document structure as determined by the layout of the text, its genre-specific content structure, and its breakdown into topical segments. Then the focus moves to phenomena of local coherence. We introduce the problem of coreference and look at methods for building chains of coreferring entities in the text. Next, the notion of coherence relation is introduced as the second important factor of local coherence. We study the role of connectives and other means of signaling such relations in text, and then return to the level of larger textual units, where tree or graph structures can be ascribed by recursively assigning coherence relations. Taken together, these descriptions can inform text summarization, information extraction, discourse-aware sentiment analysis, question answering, and the like. Table of Contents: Introduction / Large Discourse Units and Topics / Coreference Resolution / Small Discourse Units and Coherence Relations / Summary: Text Structure on Multiple Interacting Levels

Computational Methods for Corpus Annotation and Analysis Xiaofei Lu 2016-09-03 In the past few decades the use of increasingly large text corpora has grown rapidly in language and linguistics research. This was enabled by remarkable strides in natural language

processing (NLP) technology, technology that enables computers to automatically and efficiently process, annotate and analyze large amounts of spoken and written text in linguistically and/or pragmatically meaningful ways. It has become more desirable than ever before for language and linguistics researchers who use corpora in their research to gain an adequate understanding of the relevant NLP technology to take full advantage of its capabilities. This volume provides language and linguistics researchers with an accessible introduction to the state-of-the-art NLP technology that facilitates automatic annotation and analysis of large text corpora at both shallow and deep linguistic levels. The book covers a wide range of computational tools for lexical, syntactic, semantic, pragmatic and discourse analysis, together with detailed instructions on how to obtain, install and use each tool in different operating systems and platforms. The book illustrates how NLP technology has been applied in recent corpus-based language studies and suggests effective ways to better integrate such technology in future corpus linguistics research. This book provides language and linguistics researchers with a valuable reference for corpus annotation and analysis.

Linguistic Fundamentals for Natural Language Processing II Emily M. Bender 2022-06-01 Meaning is a fundamental concept in Natural Language Processing (NLP), in the tasks of both Natural Language Understanding (NLU) and Natural Language Generation (NLG). This is because the aims of these fields are to build systems that understand what people mean when they speak or write, and that can produce linguistic strings that successfully express to people the intended content. In order for NLP to scale beyond partial, task-specific solutions, researchers in these fields must be informed by what is known about how humans use language to express and understand communicative intents. The purpose of this book is to present a selection of useful information about semantics and pragmatics, as understood in linguistics, in a way that's accessible to and useful for NLP practitioners with minimal (or even no) prior training in linguistics.

Deep Learning Approaches to Text Production Shashi Narayan 2022-06-01 Text production has many applications. It is used, for instance, to generate dialogue turns from dialogue moves, verbalise the content of knowledge bases, or generate English sentences from rich linguistic representations, such as dependency trees or abstract meaning representations. Text production is also at work in text-to-text transformations such as sentence compression, sentence fusion, paraphrasing, sentence (or text) simplification, and text summarisation. This book offers an overview of the fundamentals of neural models for text production. In particular, we elaborate on three main aspects of neural approaches to text production: how sequential decoders learn to generate adequate text, how encoders learn to produce better input representations, and how neural generators account for task-specific objectives. Indeed, each text-production task raises a slightly different challenge (e.g, how to take the dialogue context into account when producing a dialogue turn, how to detect and merge relevant information when summarising a text, or how to produce a well-formed text that correctly captures the information contained in some input data in the case of data-to-text generation). We outline the constraints specific to some of these tasks and examine how existing neural models account for them. More generally, this book considers text-to-text, meaning-to-text, and data-to-text transformations. It aims to provide the audience with a basic knowledge of neural approaches to text production and a roadmap to get them started with the related work. The book is mainly targeted at researchers, graduate students, and industrials interested in text production from different forms of inputs.

Handbook of Linguistic Annotation Nancy Ide 2017-06-16 This handbook offers a thorough treatment of the science of linguistic annotation. Leaders in the field guide the reader through the process of modeling, creating an annotation language, building a corpus and evaluating it for correctness. Essential reading for both computer scientists and linguistic researchers.Linguistic annotation is an increasingly important activity in the field of computational linguistics because of its critical role in the development of language models for natural language processing applications. Part one of this book covers all phases of the linguistic annotation process, from annotation scheme design and choice of representation format through both the manual and automatic annotation process, evaluation, and iterative improvement of annotation accuracy. The second part of the book includes case studies of annotation projects across the spectrum of linguistic annotation types, including morpho-syntactic tagging, syntactic analyses, a range of semantic analyses (semantic roles, named entities, sentiment and opinion), time and event and spatial analyses, and discourse level analyses including discourse structure, co-reference, etc. Each case study addresses the various phases and processes discussed in the chapters of part one.

Natural Language Annotation for Machine Learning James Pustejovsky 2012-10-11 Create your own natural language training corpus for machine learning. Whether you’re working with English, Chinese, or any other natural language, this hands-on book guides you through a proven annotation development cycle—the process of adding metadata to your training corpus to help ML algorithms work more efficiently. You don’t need any programming or linguistics experience to get started. Using detailed examples at every step, you’ll learn how the MATTER Annotation Development Process helps you Model, Annotate, Train, Test, Evaluate, and Revise your training corpus. You also get a complete walkthrough of a real-world annotation project. Define a clear annotation goal before collecting your dataset (corpus) Learn tools for analyzing the linguistic content of your corpus Build a model and specification for your annotation project Examine the different annotation formats, from basic XML to the Linguistic Annotation Framework Create a gold standard corpus that can be used to train and test ML algorithms Select the ML algorithms that will process your annotated data Evaluate the test results and revise your annotation task Learn how to use lightweight software for annotating texts and adjudicating the annotations This book is a perfect companion to O’Reilly’s Natural Language Processing with Python.

Introduction to Linguistic Annotation and Text Analytics Graham Wilcock 2022-05-31 Linguistic annotation and text analytics are active areas of research and development, with academic conferences and industry events such as the Linguistic Annotation Workshops and the annual Text Analytics Summits. This book provides a basic introduction to both fields, and aims to show that good linguistic annotations are the essential foundation for good text analytics. After briefly reviewing the basics of XML, with practical exercises illustrating in-line and stand-off annotations, a chapter is devoted to explaining the different levels of linguistic annotations. The reader is encouraged to create example annotations using the WordFreak linguistic annotation tool. The next chapter shows how annotations can be created automatically using statistical NLP tools, and compares two sets of tools, the OpenNLP and Stanford NLP tools. The second half of the book describes different annotation formats and gives practical examples of how to interchange annotations between different formats using XSLT transformations. The two main text analytics architectures, GATE and UIMA, are then described and compared, with practical exercises showing how to configure and customize them. The final chapter is an introduction to text analytics, describing the main applications and functions including named entity recognition, coreference resolution and information extraction, with practical examples using both open source and commercial tools. Copies of the example files, scripts, and stylesheets used in the book are available from the companion website, located at the book website. Table of Contents: Working with XML / Linguistic Annotation / Using Statistical NLP Tools / Annotation Interchange / Annotation Architectures / Text Analytics

Computational Modeling of Narrative Inderjeet Mani 2022-05-31 The field of narrative (or story) understanding and generation is one of the oldest in natural language processing (NLP) and artificial intelligence (AI), which is hardly surprising, since storytelling is such a fundamental and familiar intellectual and social activity. In recent years, the demands of interactive entertainment and interest in the creation of engaging narratives with life-like characters have provided a fresh impetus to this field. This book provides an overview of the

principal problems, approaches, and challenges faced today in modeling the narrative structure of stories. The book introduces classical narratological concepts from literary theory and their mapping to computational approaches. It demonstrates how research in AI and NLP has modeled character goals, causality, and time using formalisms from planning, case-based reasoning, and temporal reasoning, and discusses fundamental limitations in such approaches. It proposes new representations for embedded narratives and fictional entities, for assessing the pace of a narrative, and offers an empirical theory of audience response. These notions are incorporated into an annotation scheme called NarrativeML. The book identifies key issues that need to be addressed, including annotation methods for long literary narratives, the representation of modality and habituality, and characterizing the goals of narrators. It also suggests a future characterized by advanced text mining of narrative structure from large-scale corpora and the development of a variety of useful authoring aids. This is the first book to provide a systematic foundation that integrates together narratology, AI, and computational linguistics. It can serve as a narratology primer for computer scientists and an elucidation of computational narratology for literary theorists. It is written in a highly accessible manner and is intended for use by a broad scientific audience that includes linguists (computational and formal semanticists), AI researchers, cognitive scientists, computer scientists, game developers, and narrative theorists. Table of Contents: List of Figures / List of Tables / Narratological Background / Characters as Intentional Agents / Time / Plot / Summary and Future Directions

Linked Lexical Knowledge Bases Iryna Gurevych 2016-07-19 This book conveys the fundamentals of Linked Lexical Knowledge Bases (LLKB) and sheds light on their different aspects from various perspectives, focusing on their construction and use in natural language processing (NLP). It characterizes a wide range of both expert-based and collaboratively constructed lexical knowledge bases. Only basic familiarity with NLP is required and this book has been written for both students and researchers in NLP and related fields who are interested in knowledge-based approaches to language analysis and their applications. Lexical Knowledge Bases (LKBs) are indispensable in many areas of natural language processing, as they encode human knowledge of language in machine readable form, and as such, they are required as a reference when machines attempt to interpret natural language in accordance with human perception. In recent years, numerous research efforts have led to the insight that to make the best use of available knowledge, the orchestrated exploitation of different LKBs is necessary. This allows us to not only extend the range of covered words and senses, but also gives us the opportunity to obtain a richer knowledge representation when a particular meaning of a word is covered in more than one resource. Examples where such an orchestrated usage of LKBs proved beneficial include word sense disambiguation, semantic role labeling, semantic parsing, and text classification. This book presents different kinds of automatic, manual, and collaborative linkings between LKBs. A special chapter is devoted to the linking algorithms employing text-based, graph-based, and joint modeling methods. Following this, it presents a set of higher-level NLP tasks and algorithms, effectively utilizing the knowledge in LLKBs. Among them, you will find advanced methods, e.g., distant supervision, or continuous vector space models of knowledge bases (KB), that have become widely used at the time of this book's writing. Finally, multilingual applications of LLKB's, such as cross-lingual semantic relatedness and computer-aided translation are discussed, as well as tools and interfaces for exploring LLKBs, followed by conclusions and future research directions.

Introducing Electronic Text Analysis Svenja Adolphs 2006-09-27 Introducing Electronic Text Analysis is a practical and much needed introduction to corpora – bodies of linguistic data. Written specifically for students studying this topic for the first time, the book begins with a discussion of the underlying principles of electronic text analysis. It then examines how these corpora enhance our understanding of literary and non-literary works. In the first section the author introduces the concepts of concordance and lexical frequency, concepts which are then applied to a range of areas of language study. Key areas examined are the use of on-line corpora to complement traditional stylistic analysis, and the ways in which methods such as concordance and frequency counts can reveal a particular ideology within a text. Presenting an accessible and thorough understanding of the underlying principles of electronic text analysis, the book contains abundant illustrative examples and a glossary with definitions of main concepts. It will also be supported by a companion website with links to on-line corpora so that students can apply their knowledge to further study. The accompanying website to this book can be found at http://www.routledge.com/textbooks/0415320216

Pretrained Transformers for Text Ranking Jimmy Lin 2022-06-01 The goal of text ranking is to generate an ordered list of texts retrieved from a corpus in response to a query. Although the most common formulation of text ranking is search, instances of the task can also be found in many natural language processing (NLP) applications.This book provides an overview of text ranking with neural network architectures known as transformers, of which BERT (Bidirectional Encoder Representations from Transformers) is the best-known example. The combination of transformers and self-supervised pretraining has been responsible for a paradigm shift in NLP, information retrieval (IR), and beyond. This book provides a synthesis of existing work as a single point of entry for practitioners who wish to gain a better understanding of how to apply transformers to text ranking problems and researchers who wish to pursue work in this area. It covers a wide range of modern techniques, grouped into two high-level categories: transformer models that perform reranking in multi-stage architectures and dense retrieval techniques that perform ranking directly. Two themes pervade the book: techniques for handling long documents, beyond typical sentence-by-sentence processing in NLP, and techniques for addressing the tradeoff between effectiveness (i.e., result quality) and efficiency (e.g., query latency, model and index size). Although transformer architectures and pretraining techniques are recent innovations, many aspects of how they are applied to text ranking are relatively well understood and represent mature techniques. However, there remain many open research questions, and thus in addition to laying out the foundations of pretrained transformers for text ranking, this book also attempts to prognosticate where the field is heading.

Grammatical Inference for Computational Linguistics Jeffrey Heinz 2022-06-01 This book provides a thorough introduction to the subfield of theoretical computer science known as grammatical inference from a computational linguistic perspective. Grammatical inference provides principled methods for developing computationally sound algorithms that learn structure from strings of symbols. The relationship to computational linguistics is natural because many research problems in computational linguistics are learning problems on words, phrases, and sentences: What algorithm can take as input some finite amount of data (for instance a corpus, annotated or otherwise) and output a system that behaves "correctly" on specific tasks? Throughout the text, the key concepts of grammatical inference are interleaved with illustrative examples drawn from problems in computational linguistics. Special attention is paid to the notion of "learning bias." In the context of computational linguistics, such bias can be thought to reflect common (ideally universal) properties of natural languages. This bias can be incorporated either by identifying a learnable class of languages which contains the language to be learned or by using particular strategies for optimizing parameter values. Examples are drawn largely from two linguistic domains (phonology and syntax) which span major regions of the Chomsky Hierarchy (from regular to context-sensitive classes). The conclusion summarizes the major lessons and open questions that grammatical inference brings to computational linguistics. Table of Contents: List of Figures / List of Tables / Preface / Studying Learning / Formal Learning / Learning Regular Languages / Learning Non-Regular Languages / Lessons Learned and Open Problems / Bibliography / Author Biographies

Linguistic Modeling of Information and Markup Languages Andreas Witt 2010-01-09 This book covers recent developments in the field,

from multi-layered mark-up and standards to theoretical formalisms to applications. It presents results from international research in text technology, computational linguistics, hypertext modeling and more.

A Practical Handbook of Corpus Linguistics Magali Paquot 2021-05-04 This handbook is a comprehensive practical resource on corpus linguistics. It features a range of basic and advanced approaches, methods and techniques in corpus linguistics, from corpus compilation principles to quantitative data analyses. The Handbook is organized in six Parts. Parts I to III feature chapters that discuss key issues and the know-how related to various topics around corpus design, methods and corpus types. Parts IV-V aim to offer a user-friendly introduction to the quantitative analysis of corpus data: for each statistical technique discussed, chapters provide a practical guide with R and come with supplementary online material. Part VI focuses on how to write a corpus linguistic paper and how to meta-analyze corpus linguistic research. The volume can serve as a course book as well as for individual study. It will be an essential reading for students of corpus linguistics as well as experienced researchers who want to expand their knowledge of the field.

Automatic Text Simplification Horacio Saggion 2022-05-31 Thanks to the availability of texts on the Web in recent years, increased knowledge and information have been made available to broader audiences. However, the way in which a text is written—its vocabulary, its syntax—can be difficult to read and understand for many people, especially those with poor literacy, cognitive or linguistic impairment, or those with limited knowledge of the language of the text. Texts containing uncommon words or long and complicated sentences can be difficult to read and understand by people as well as difficult to analyze by machines. Automatic text simplification is the process of transforming a text into another text which, ideally conveying the same message, will be easier to read and understand by a broader audience. The process usually involves the replacement of difficult or unknown phrases with simpler equivalents and the transformation of long and syntactically complex sentences into shorter and less complex ones. Automatic text simplification, a research topic which started 20 years ago, now has taken on a central role in natural language processing research not only because of the interesting challenges it posesses but also because of its social implications. This book presents past and current research in text simplification, exploring key issues including automatic readability assessment, lexical simplification, and syntactic simplification. It also provides a detailed account of machine learning techniques currently used in simplification, describes full systems designed for specific languages and target audiences, and offers available resources for research and development together with text simplification evaluation techniques.