Natural Language Processing For Historical Texts (Synthesis Lectures On Human Language Technologies)
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

Book Information

Series: Synthesis Lectures on Human Language Technologies
Paperback: 158 pages
Publisher: Morgan & Claypool Publishers (September 28, 2012)
Language: English
ISBN-10: 1608459462
Product Dimensions: 7.5 x 0.4 x 9.2 inches
Shipping Weight: 13 ounces (View shipping rates and policies)