When Printed Hypertexts Go Digital: Information Extraction from the Parsing of Indices

Matteo Romanello
The Perseus Project
Tufts University
Medford, MA
matteo.romanello@tufts.edu

Monica Berti
The Perseus Project
Tufts University
Medford, MA
monica.berti@tufts.edu

Alison Babeu
The Perseus Project
Tufts University
Medford, MA
alison.jones@tufts.edu

Gregory Crane
The Perseus Project
Tufts University
Medford, MA
gregory.crane@tufts.edu

ABSTRACT

Modern critical editions of ancient works generally include manually created indices of other sources quoted in the text. Since indices can be considered as a form of domain specific language, the paper presents a parsing-based approach to the problem of extracting information from them to support the creation of a collection of fragmentary texts. This paper first considers the characteristics and structure of quotation indices and their importance when dealing with fragmentary texts. It then presents the results of applying a fuzzy parser to the OCR transcription of an index of quotations to extract information from potentially noisy input.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: [Hypertext/Hypermedia]

General Terms

Design, Experimentation.

Keywords

Printed hypertexts, indices, information extraction, parsing.

1. INTRODUCTION

In recent years, mass digitization initiatives have made accessible the page images of an increasing number of modern editions. Now we can access not only the text but also the paratextual apparatus of each digital edition, namely prefaces, notes, critical apparatuses and indices. The ongoing work presented in this paper is related to a project which aims to provide the Perseus Digital Library with a collection of fragmentary texts, specifically a collection of historical Greek fragments. The topic of converting printed scholarly materials to digital hypertexts has a long research history [7, 1] including recent inquiries into the hypertextual nature of historical publications [3], [4]. This paper proposes the automatic parsing of manually created *indices scriptorum* (i.e. indices of quotations) as an approach to reuse the efforts made over decades by scholars on individuating and indexing citations inside texts in order to create new digital tools. Specifically we give an example of this approach by showing how it is possible to use information extracted from parsing the indices of works containing witnesses of fragments to also support the automatic identification and markup of those fragments in the text.

HT'09, June 29–July 1, 2009, Torino, Italy.
ACM 978-1-60558-486-7/09/06.

2. INDICES OF QUOTATIONS

The indices of quotations found in many modern critical editions of classical authors can be thought of as the hypertext through which an editor creates internal links to those passages in their edited work that contain quotations from other ancient sources. These indices also provide outward links to the entire body of classical literature by listing quotations of other surviving works. Indices are worth parsing since we can reconstruct internal and external links between different texts and we can extract information such as lists of names, epithets of authors, titles of works, canonical citations used by scholars, and the variants and conjectures reported by the editor.

When considering fragmentary texts such indices assume a particular importance. Indeed fragments are a straightforward example of how quotations can become a crucial factor in the survival of a literary text, since they are basically passages of works that only survived because they were quoted within surviving works by other authors. Since fragments are essentially quotations, the indices of quotations in modern editions of texts containing witnesses of fragments can serve as an essential source of information about them.

3. INDEX PARSING AND INFORMATION EXTRACTION

The main assumption for building a parser of printed indices is that an index constitutes a domain-specific language and that the syntactic disposition of its lexical components is subject to a grammar of rules that can be preliminarily defined. [2] recently demonstrated how another kind of scholarly paratext contained within critical editions of classical texts, namely the critical apparatus, is characterized by

such a formal, consistent and unambiguous structure that it allows us to use parsing techniques to extract information from it.

An index of quotations is mainly structured as entries where each entry corresponds to an author and where the name of the author constitutes the so-called lemma, namely the headword. For each author the editor lists every work of that same author that is cited in the text. After the lexical elements of the index were identified, a grammar of syntactic rules was specified, defining for each of them the corresponding sequence of tokens to be matched. This grammar expressed in the Extended Backus-Naur Form (EBNF) was then transformed into a Java parser by using the ANTLR parser generator [6].

The errors contained in the OCR transcription of the index, however, introduce a component of uncertainty in the parsing that needs to be properly considered. A deterministic grammar implying strict conformity to syntactic rules is not fully suitable since it cannot adapt flexibly to OCR errors. Instead such an issue can be properly handled by using a fuzzy parsing approach [5] since it allows us to isolate noisy data and to detect only meaningful sequences of lexical elements. Since the knowledge domain of indices is limited and the semantic content of sequences can thus be identified by grammar rules, it becomes possible to correct some of the errors due to the OCR with some benefits in terms of scalability of the proposed approach. The parsing results are improved by recovering basically from two kinds of errors: 1) errors (handled by the parser as exceptions) due to unrecognized sequences of tokens; 2) imprecise content of tokens (e.g. misspelled names)

The proposed system has been prototyped on the text of Athenaeus' *Deipnosophistae*. Our knowledge of some ancient works relies uniquely upon this text as a witness of fragments. The considered index of quotations is drawn from the Kaibel edition of the *Deipnosophistae* where it spans over 111 pages containing references to quotations attributed to 786 authors. The developed framework consists of 1) a TEI XML edition of the text that the index is referring to (this edition is based on the text published by G. Kaibel in 1887-1890); 2) an OCR transcription of the index to be parsed[1] with an accuracy of 96.73% which was produced by using an OCRopus installation trained on Ancient Greek[2]; 3) a set of external knowledge sources (such as name lists) that can be used for the error correction

The result of parsing the index of quotations is a machine actionable tree representing its hierarchical structure, where each author and work quoted in the text is associated to a reference indicating its position in the text. Starting from this resulting tree it is possible to automatically tag the quotations in the text and thus to reconstruct the hypertextual links between the index and the text. Heuristics based on elements of the paratext such as quotation marks, indentations and canonical references are then applied to determine the boundaries of text quotations with more precision. The output of this phase is a collection of links to the precise position of each quotation in the text expressed as a pair of pointers to the words that serve as the boundaries of the quoted passage.

[1]The OCR was performed on the edition available at `http://books.google.com/books?id=kB1BAAAAMAAJ`

[2]`http://sites.google.com/site/ocropus/languages/ancient-greek`

In addition to the identification and markup of quotations, using this approach makes it possible to extract from the index and then identify in the text other information about quoted authors and works. This information includes Greek titles of works, the Greek names and epithets of their authors, as well as the periphrastic expressions used to refer to them. Much of this information is not currently marked up within digital editions, and furthermore the available catalogs of metadata about ancient authors and works typically only include normalized forms of names in Latin or in modern languages.

4. CONCLUSIONS AND FUTURE WORK

Even though parsing techniques are widely known and used in other disciplines, the recent availability of a large number of digitized modern editions provides scholars in classics with essential materials that have never before been leveraged by using a parsing-based approach. This paper has shown how a certain amount of information can be easily extracted even from the noisy OCR transcription of printed indices. Extracted information can be reused to train a new generation of more sophisticated tools for information extraction and further text analysis can be then applied to the automatically identified text quotations.

5. ACKNOWLEDGMENTS

Grants from the Andrew W. Mellon Foundation ("The Cyberedition Project") and the NEH in conjunction with the IMLS ("Scalable Named Entity Identification in Classical Studies") provided support for this work. We also gratefully acknowledge Federico Boschetti of CIMeC-University of Trento (Italy) for his useful suggestions.

6. REFERENCES

[1] A. Belaïd, I. Turcan, J. M. Pierrel, Y. Belaïd, Y. Hadjamar, and H. Hadjamar. Automatic indexing and reformulation of ancient dictionaries. In *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*, page 342. IEEE Computer Society, 2004.

[2] F. Boschetti. Methods to extend greek and latin corpora with variants and conjectures: Mapping critical apparatuses onto reference text. In *Proceedings of the Corpus Linguistics Conference (CL2007)*, 2007.

[3] O. Kolak and B. N. Schilit. Generating links by mining quotations. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 117–126, Pittsburgh, PA, USA, 2008. ACM.

[4] D. Kolb. Scholarly hypertext: self-represented complexity. In *Proceedings of the eighth ACM conference on Hypertext*, pages 29–37, Southampton, United Kingdom, 1997. ACM.

[5] R. Koppler. A systematic approach to fuzzy parsing. *Software Practice and Experience*, 27:637—649, 1997.

[6] T. J. Parr and R. W. Quong. ANTLR: a predicated-LL(k) parser generator. *Software Practice and Experience*, 25:789—810, 1995.

[7] D. R. Raymond and F. W. Tompa. Hypertext and the new oxford english dictionary. In *Proceedings of the ACM conference on Hypertext*, pages 143–153, Chapel Hill, North Carolina, United States, 1987. ACM.