Cristina Bosco, Sara Tonelli and Fabio Massimo Zanzotto (dir.)

Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015
3-4 December 2015, Trento

Accademia University Press

On Mining Citations to Primary and Secondary Sources in Historiography

Giovanni Colavizza and Frédéric Kaplan

DOI: 10.4000/books.aaccademia.1439
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2015
Published on OpenEdition Books: 11 November 2016
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899200008

On Mining Citations to Primary and Secondary Sources in Historiography

Giovanni Colavizza, Frédéric Kaplan
EPFL, CDH, DH Laboratory, Lausanne, Switzerland
{giovanni.colavizza,frederic.kaplan}@epfl.ch

Abstract

English. We present preliminary results from the Linked Books project, which aims at analysing citations from the historiography on Venice. A preliminary goal is to extract and parse citations from any location in the text, especially footnotes, both to primary and secondary sources. We detail a pipeline for these tasks based on a set of classifiers, and test it on the *Archivio Veneto*, a journal in the domain.

Italiano. *Presentiamo i primi risultati del progetto Linked Books, per l'analisi delle citazioni della storiografia su Venezia. Ci prefiggiamo l'estrazione e l'analisi delle citazioni da ogni posizione nei testi, specialmente note a pi pagina, sia a fonti primarie che secondarie. Discutiamo una serie di classificatori con questo obiettivo, valutandone i risultati su Archivio Veneto, una rivista del settore.*

1 Introduction

The Linked Books project is part of the Venice Time Machine[1], a joint effort to digitise and study the history of Venice by digital means. The project goal is to analyse the history of Venice through the lens of citations, by network analytic methods. Such research is interesting because it could unlock the potential of the rich semantics of the use of citations in humanities. A preliminary step is the extraction and normalization of citations, which is a challenge in itself. In this paper we present the first results on this last topic, over a corpus of journals and monographs on the history of Venice, digitised in partnership with the Ca' Foscari Humanities Library and the Marciana Library.

[1] http://vtm.epfl.ch/.

Our contribution is three-fold. First, we address the problem of extracting citations in historiography, something rarely attempted before. Secondly, we extract citation from footnotes, with plain text as input. Lastly, we deal at the same time with two different kind of citations: to primary and to secondary sources. A primary source is a documentary evidence used to support a claim, a secondary source is a scholarly publication (Wiberley Jr, 2010). In order to solve this problem, we propose a pipeline of classifiers dealing with citation detection, extraction and parsing.

The paper is organised as follows: a state of the art in Section 2 is followed by a methodological section explaining the pipeline and applied computational tools. A section on experiments follows, conclusions and future steps close the paper.

2 Related work

Sciences have largely used quantitative citation data to study their practices, whilst humanities remained largely outside of the process (Ardanuy, 2013). Difficulties of a concrete nature along with peculiar features of humanistic discourse make the task not trivial.

The lack of citation data for the humanities is well recognised, both for monographs and other kind of secondary literature (Heinzkill, 1980; Larivière et al., 2006; Linmans, 2009; Hammarfelt, 2011; Sula and Miller, 2014). Furthermore, citations are deployed within humanities in multifaceted ways, posing further challenges to their extraction and understanding (Grafton, 1999; Hellqvist, 2009; Sula and Miller, 2014).

One core element of citations in humanities, and especially so History, is the distinction between primary and secondary sources, and the quantitative and qualitative importance of both (Frost, 1979; Hellqvist, 2009). Little previous work on the use of primary sources via citations exist, with few exceptions in the domains of biblical stud-

ies and Classics (Murai and Tokosumi, 2008; Romanello, 2014).

The literature on citation extraction mirrors this scenario. As far as the citations to secondary sources are concerned, the development of automatic citation indexing systems has been a well explored area of research over the last two decades, starting from the seminal work of Giles et al. (1998). Increasingly, researchers are also tackling the problem of locating citations within the structure of documents (Lopez, 2009; Kim et al., 2012b; Heckmann et al., 2014). The extraction of citations to primary sources is instead a largely unexplored area, where recent effort has been produced within the fields of Classics (Romanello et al., 2009; Romanello, 2013; Romanello, 2014) and law (Francesconi et al., 2010; Galibert et al., 2010).

3 Approach

We propose a three-staged incremental pipeline including the following steps:

1. **Text block detection** of contiguous lines of text likely to contain citations, usually footnotes. The motivation for this preliminary step, inspired by Kim et al. (2012b), is to individuate the footnote space of a publication, as footnotes can span multiple pages.

2. **Citation extraction** within their boundaries over one or more contiguous text lines. This stage entails a token by token classification. A further sub-step is the classification of a citation as being *Primary* or *Secondary*, meaning to primary or secondary sources respectively.

3. **Citation parsing**, token by token, to detect all relevant components over a set of 50 mutually exclusive classes (e.g. *Author*, *Title* and *PublicationDate* for citations to secondary sources, or *Archive*, *Fond* and *Series* for primary sources).

The first step is dealt with using a SVM classifier,[2] initially trained with a small set of morphological features.

The second and last steps are approached with a group of CRF classifiers trained over a rich set of features, considering a bi-gram and tri-gram context, both backwards and forward. We train the models with Stochastic Gradient Descent and L2 regularisation, using the CRFSuite and default parameters (Okazaki, 2007).

Conditional Random Fields and Supporting Vector Machines are state-of-the-art models in the field of citation extraction since the work of Peng and McCallum (2006), and were introduced first by Cortes and Vapnik (1995) and Lafferty et al. (2001) respectively.

4 Experiments

The corpus is first digitised,[3] then OCRed using a commercial product with no extra training.[4] Our tests are based on an annotated sample of pages from the *Archivio Veneto*—a scholarly journal in Italian specialised in the History of Venice—randomly selected from a corpus of 92 issues from the year 1969 to 2013. The sample consists of 1138 annotated pages, for a total of 6257 annotated citations. Proper evaluation of the OCR quality and inter-annotator agreement are still pending at this stage. The annotation phase has been carried out with Brat.[5] No text format features—i.e. italics or type module—are used for the moment, and will be considered in a subsequent phase of the project.

4.1 Text block detection

The first classification step is a boolean one, where we are interested in knowing if a line of text, or a group of contiguous lines, is likely to contain citations, therefore likely to be a footnote. Text blocks are defined as groups of k contiguous lines of text. This step is required by the nature of footnotes, which can span over multiple pages demanding their proper identification in order to define the input space for subsequent stages in the pipeline. For each block we extract the following features: 1- **General**: line number (to detect footnotes); 2- **Morphological**[6]: punctuation frequency, frequency of digits, frequency of uppercase and lower-case characters, number of white spaces, number of characters, frequency of abbreviations according to multiple patterns, average word length, average frequency of specific punctuation symbols (“.”, “,”, “(”, “)”, “[”, “]”); 3- **Boolean**: if the chunk begins with a possible

[2] Using Python sklearn package.

[3] With 4DigitalBooks DLmini scanners.

[4] Abbyy FineReader Corporate 12.

[5] `http://brat.nlplab.org/`.

[6] Frequencies are always assessed character by character.

acronym or with a digit. After experimental tuning, we settle for a poly-linear model of degree 2 over a set of alternatives (degrees 1 to 10), which has the added value of maximizing recall, the most important metric at this early stage. The best division into text-blocks is found to be with $k = 2$. The evaluation of this step, based on a randomly-selected third of the annotated data (3633 blocks, 2204 negative and 1429 positive), is reported in Table 1.

Task	Precision	Recall	F1-score
no-citation	0.96	0.95	0.96
citation	0.92	0.95	0.93
avg / total	0.95	0.95	0.95

Table 1: Evaluation results for Text block detection.

Our results compare with others applying similar filtering methods (Kim et al., 2012a). In the future we will test a confidence classification with threshold lower than 0.5, as to further improve recall over precision.

4.2 Citation extraction

Given a text block likely to contain citations, we address the problem of citation extraction, meaning tokenizing the block and tagging each token as being part of a citation or not. For this phase and the next, text blocks are merged as to avoid any input being considered twice or more in the training and test sets. We merge together contiguous text lines likely to contain a citation, and consider k extra context (lines of text without citations) before and after. The set of features used for this step is organised in the following classes:[7]

1. **Shape of the token**: according to each character being upper-case, lower-case or punctuation. E.g. "UUU." for a token of length 4 with 3 upper-case characters and a final dot.

2. **Type of the token**: according to a set of classes such as if the token is a digit, or made of all upper-case letters, etc.

3. **Boolean features**: if the token is a 2 or 4 digit number, if it contains digits, if it contains upper or lower case characters, etc.

4. **Other features**: the token itself and it's position in the current line.

A more limited set of features is also considered in a bi and tri-gram conditioning over a sliding window within the preceding and following 3 tokens, namely: the tokens themselves, their shape and type, their position in the line.

The evaluation was conducted on a set of 19852 tokens (5240 primary and 14612 secondary) and 1056 text blocks, corresponding to a random third of the annotated corpus. The most balanced context turned out to be $k = 2$, results in Table 2. The performance is acceptably high in terms of overall item accuracy (0.95). In general, a higher context k means trading off precision for recall. Instance accuracy is apparently much lower (0.504), we must however remember that an instance at this level is a text block, possibly containing several non contiguous citations. Instance accuracy at the citation level improves to 0.78, and 0.84 if we tolerate for 1 token of difference between the golden standard and automatic tagging of a citation. We therefore attain results comparable to those Lopez (2010) got for the task of individuating non-patent references in patent text bodies.

Task	Precision	Recall	F1-score
no-citation	0.978	0.917	0.947
citation	0.926	0.98	0.953
avg / total	0.952	0.949	0.95

Table 2: Evaluation results for Citation extraction.

We further explored if a classifier trained with the same features could properly distinguish citations to primary and secondary sources. For this task each citation is parsed independently, assuming proper segmentation from the previous step. We attain an overall item accuracy of 0.967 and instance accuracy of 0.928 over the same training and testing sets. The fact that this classifier performs well allows us to consider the macro-category (primary or secondary) as a feature in the parsing step. Results in Table 3.

4.3 Citation parsing

This step involves the parsing of an extracted citation in order to individuate its components. The same set of features as before is used for each token, with the addition of:

- **Enhanced boolean features**: if the token is

[7] The full list of features is available upon request and partially inspired by Okazaki (2007).

Task	Precision	Recall	F1-score
primary	0.968	0.904	0.935
secondary	0.966	0.989	0.978
avg / total	0.967	0.947	0.956

Table 3: Evaluation results for Primary and Secondary Citation classification.

a time span (e.g. "1600-1700"), if it might be a Roman number, or an abbreviation.

- **The macro-category** (primary or secondary), as an indicator of the typology of the citation.

Task	Precision	Recall	F1-score
Author	0.939	0.958	0.948
Title	0.873	0.989	0.928
Pub.Place	0.927	0.899	0.913
Pub.Year	0.927	0.861	0.893
Pagination	0.961	0.978	0.969
Archive	0.968	0.912	0.939
ArchivalRef.	0.909	0.884	0.896
Folder	0.955	0.938	0.947
Registry	0.957	0.901	0.928
Cartulation	0.938	0.908	0.921
Foliation	0.862	0.890	0.875

Table 4: Evaluation results for Citation parsing: without macro-category feature.

Task	Precision	Recall	F1-score
Author	0.94	0.957	0.948
Title	0.9	0.984	0.94
Pub.Place	0.931	0.908	0.919
Pub.Year	0.945	0.893	0.918
Pagination	0.953	0.984	0.968
Archive	0.969	0.919	0.943
ArchivalRef.	0.901	0.895	0.898
Folder	0.956	0.942	0.949
Registry	0.971	0.901	0.935
Cartulation	0.964	0.934	0.949
Foliation	0.892	0.884	0.888

Table 5: Evaluation results for Citation parsing: with macro-category feature.

We test over a random 30% of the corpus and report only results of parsing with no extra context, which predictably gave the best results. Overall item and instance accuracy are 0.884 and 0.575 without the macro-category feature, and 0.893 and 0.592 with it. The testing set is comparable and proportional in size, yet different in sampling to the one used in step 2. Results in Table 4 and Table 5 only report the most significant classes in order to understand a citation, for citations secondary (above) and primary sources (below) respectively.[8]

The macro-category has only a marginal, albeit positive impact. Furthermore, some categories are either under-represented in terms of training instances, or easily mistaken for another one, contributing to the overall degradation of results. Such is the case for *Editor* or *Curator*, frequently classified as *Author*. In general several categories could be grouped, and lookup features—over list of names or library catalogues—should greatly improve our results.

The model performs well for the most significant categories, in comparison to models trained on more data and/or fewer categories and/or on references and not footnote citations. Specifically, we improve on Lopez (2010), Kim et al. (2012b), Romanello (2013), and compare to Heckmann et al. (2014).

5 Conclusions and future work

We presented a pipeline for recognizing and parsing citations to primary and secondary sources from historiography on Venice, with a case study on the *Archivio Veneto* journal. A first filtering step allows us to detect text blocks likely to contain citations, usually footnotes, by a SVM classifier trained on a simple set of morphological features. We then detect citation boundaries and macro-categories (to primary and secondary sources) using more rich features and CRFs. The last step in our pipeline is the fine-grained parsing of each extracted citation, in order to prepare them for further processing and analysis.

In the future we plan to design more advanced feature sets, first of all considering text format features. Secondly, we will implement the next package of our chain: an error-tolerant normalizer which will uniform all citations to the same primary or secondary source within a publication, as a means to minimise the impact of classification errors during previous steps.

[8] The full list of results is available upon request.

Acknowledgments

We thank Maud Ehrmann and Jean-Cédric Chappelier, EPFL, for useful comments.

The project is funded by the Swiss National Fund under Division II, project number 205121_159961.

References

Jordi Ardanuy. 2013. Sixty years of citation analysis studies in the humanities (1951-2010). *Journal of the American Society for Information Science and Technology*, 64(8):1751–1755.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. Semantic processing of legal texts - where the language of law meets the law of language.

Carolyn O. Frost. 1979. The use of citations in literary research: A preliminary classification of citation functions. *The Library Quarterly*, pages 399–414.

Olivier Galibert, Sophie Rosset, Xavier Tannier, and Fanny Grandry. 2010. Hybrid citation extraction from patents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 530–534.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98.

Anthony Grafton. 1999. *The Footnote: a Curious History*. Harvard University Press.

Björn Hammarfelt. 2011. Interdisciplinarity and the intellectual base of literature studies: citation analysis of highly cited monographs. *Scientometrics*, 86(3):705–725.

D. Heckmann, A. Frank, M. Arnold, P. Gietz, and C. Roth. 2014. Citation segmentation from sparse and noisy data: a joint inference approach with Markov logic networks. *Digital Scholarship in the Humanities*.

Richard Heinzkill. 1980. Characteristics of references in selected scholarly english literary journals. *The Library Quarterly*, pages 352–365.

Björn Hellqvist. 2009. Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology*, 61(2):310–318.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2012a. Annotated bibliographical reference corpora in Digital Humanities. In *Language Resources and Evaluation Conference*, pages 494–501.

Young-Min Kim, Patrice Bellot, Elodie Faath, and Marin Dacos. 2012b. Automatic annotation of incomplete and scattered bibliographical references in Digital Humanities papers. In *Conférence en Recherche de Information et Applications*, pages 329–340.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Vincent Larivière, Yves Gingras, and Éric Archambault. 2006. Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics*, 68(3):519–533.

A. J. M. Linmans. 2009. Why with bibliometrics the humanities does not need to be the weakest link: Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics*, 83(2):337–354.

Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries*, pages 473–474.

Patrice Lopez. 2010. Automatic extraction and resolution of bibliographical references in patent documents. In *Advances in Multidisciplinary Retrieval*, pages 120–135.

Hajime Murai and Akifumi Tokosumi. 2008. Extracting concepts from religious knowledge resources and constructing classic analysis systems. In *Large-Scale Knowledge Resources. Construction and Application*, pages 51–58.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using Conditional Random Fields. *Information Processing & Management*, 42(4):963–979.

Matteo Romanello, Federico Boschetti, and Gregory Crane. 2009. Citations in the digital library of Classics: extracting canonical references by using Conditional Random Fields. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 80–87.

Matteo Romanello. 2013. Creating an annotated corpus for extracting canonical citations from classics-related texts by using active annotation. In *Computational Linguistics and Intelligent Text Processing*, volume 7816, pages 60–76.

Matteo Romanello. 2014. Mining citations, linking texts. *Institute for the Study of the Ancient World Papers 7.24*.

Chris A. Sula and Matt Miller. 2014. Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3):452–464.

Stephen E. Wiberley Jr. 2010. Humanities literatures and their users. In *Encyclopedia of Library and Information Sciences*, pages 2197–2204.