Identifying poetry based on library catalogue metadata, Hege Roivainen

Changes in printing reflect historical turning points: what has been printed, when, where and by whom are all derivatives of contemporary events and situations. Excessive need for war propaganda brings out more pamphlets from the printing presses; university towns produce dissertations, which scientific development can be deduced from; and strict oppression and censorship might allow only religious publications by government-approved publishers. National library metadata catalogues have been used as sources for studying these turning points.

The traditional way of using national library metadata catalogues in research is simply for finding the location of physical reference books in the library (Altick & Fenstermaker, 1992, 155-182). This kind of research is qualitative in nature - based on close reading and requiring a profound knowledge of the subject matter. While library catalogues may be exploited in this manner for first appearances of various phenomena, the extent of new innovations will not be verified. For example, close readings do not reveal, at least easily, the timeline of Luther's publications, or what portion of books actually were octavo-sized, and when the increase in this format occurred. National library catalogues often contain, more or less, complete records of practically everything published in a certain country or linguistic area in a certain time period. Metadata included often covers information about: physical properties, such as book size and page counts; publisher details; publication places; and so forth. This has made them ideal sources for researchers interested in quantitative analysis and computational approaches aimed at connecting historical turning points and measurable changes in printing. For example, the impact of a new concept can be measured against the amount of re-publications, or the spread of the book, which introduced a new idea (Lahti, Ilomäki & Tolonen, 2015). What is more, linking library metadata to the full text of the books has made it possible to analyse the changes in the usage of words in massive corpora, while still limiting analysis to relevant books (Kanner et al., 2017).

In all these cases, however, computational methods work better the more complete the corpus is, and in the case of library catalogues, there are often deficiencies in annotations. Tthe reasons for this are varied: annotating resources might have been limited, or the annotation rules may have varied between different libraries in cases where catalogues have been amalgamated, or rules could have simply changed. (Karian, 2011).

One area which could be particularly important for researchers is genre. The genre field could be used to restrict the corpus to contain every one of the books that are needed and nothing more. From this subset, then, there would be the possibility of drawing timelines or graphs based on bibliographic metadata, or in the case of full texts, the language or contents of a complete corpus could be analysed. Yet, despite the potential significance of genre information, the field is often unannotated.

My research is a case study which aims at identifying works of poetry in the English Short Title Catalogue (ESTC)[1]. Poetry was chosen, first, because it is a fairly common genre in the ESTC, and second, it is a point of interest for literary researchers. A nearly complete subset of English poetry would allow for large-scale quantitative poetry analysis. With regard to the ESTC: the catalogue contains nearly half a million records of books printed either in English or in Great Britain in the early modern era. Genre information, however, only exists for approximately one fourth of the records.[2]

[1] I had a data dump of ESTC at my disposal for research purposes, so I did not use the online version (http://estc.bl.uk/).
[2] Each record in the ESTC may have multiple genre definitions which are based on the Rare Books and Manuscripts Section (RBMS) genre descriptions. (RBMS, 1991). Also, the depth of categorization varies from micro- to macro-level, which makes assembling relevant subsets difficult.

Rather than relying solely on annotations, a model for machine learning can be taught that uses the more complete aspects of the records, and then genre information deduced from it. In this case, my solution for handling incomplete genre information was turning the genre detection task into a binary classification problem. Each genre value in the catalogue genre labels is judged either as belonging to poetry or non-poetry based on "family resemblance."[3] To this end, I used existing RBMS genre descriptions focusing on versified text, excluding music or performance focused genres such as 'Opera' and 'Plays', but including 'Songs' and 'Ballads'[4]. In the ESTC there are almost 14,000 records which have exactly 'Poems' as a genre, but over 35,000 records containing a genre label describing poetry in this manner. I then took a subset of all the records where the genre field was annotated, and divided it further into training and testing sets. I then trained and tested several different models, which were mostly extracted from the title and subtitle fields. Each of the models was based on a different aspect of the records (such as bag-of-words of the most common words in poetry book titles, part-of-speech tags, or topics). From these models I created a superset of best performing features, which I again tested with the same material. The resulting model performed well: despite the compactness of the metadata, poetry books could be tracked from the test data with a precision over 95%.

I then used the superset of features to seek poetry books without the annotated genre field. This resulted in identifying 13,000 unannotated poetry books from nearly 340,000 records without genre annotation. A sample of one hundred of both poetry and non-poetry findings to manually estimate the correctness of the predictions showed precision of 73%. These results were hampered by an annotation bias in the catalogue. The bias seems to come from two factors: first, one-paged broadside poems have largely had their genre correctly identified in the metadata. Secondly, these works of poetry have been identified as two-page works. Therefore, there is a statistical weighting issue in the training material which identifies two-page works as poetry. In addition, when dealing with works longer than a page, poetry books seem to be annotated more comprehensively than non-poetry. Therefore, by excluding broadsheets from my samples precision rose to 98%.

My research strongly suggest that semi-supervised learning can be applied to library catalogues to fill in missing annotations, but this requires close attention to avoid possible pitfalls.

References

Richard D. Altick and John J. Fenstermaker. *The Art of Literary Research*. Fourth Edition. Norton ; New York. 1993. ISBN 0-393-96240-7.

ESTC. English Short Title Catalogue. URL http://estc.bl.uk/. Accessed 2018-02-05.

Alastair Fowler. *Kinds of Literature. An Introduction to the Theory of Genres and Modes.* Clarendon Press ; Oxford. 1982. ISBN 0-19-812812-6.

John Frow. *Genre*. The New Critical Idiom. Routledge London ; New York, 2006. ISBN 0-415-28063-X.

[3] In genre theory, the concept of family resemblance is often applied to define genres (for example Wolf, 2015; Fowler, 1982, 40-44). The basic idea behind family resemblance is that not all the genre markers have to be present for a text to be considered as belonging to the genre. This is especially emphasized in library catalogues, where the full texts are not available.

[4] There exists a conventionalized main genre division between prose, drama and lyric from the 17th century (Frow, 2006, 59). I have adapted this division.

Antti Kanner, Jani Marjanen, Ville Vaara, Hege Roivainen, Viivi Lähteenoja, Laura Tarkka-Robinson, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. OCTAVO – Analysing Early Modern Public Communication [poster]. Presented in Digital Humanities at Oxford Summer School. 2017. URL https://comhis.github.io/posters/octavo/. Accessed 2018-02-05.

Stephen Karian. The limitations and possibilities of the estc. *The age of Johnson*, 21:283–297, 2011.

Leo Lahti, Niko Ilomäki, and Mikko Tolonen. A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800. *LIBER Quarterly*, 25(2):87, Dec 2015. ISSN 2213-056X. doi: https://doi.org/10.18352/lq.10112.

RBMS. Genre Terms : A Thesaurus for Use in Rare Book and Special Collections Cataloguing, 1991. URL https://rbms.info/vocabularies/genre/alphabetical_list.htm. Accessed 2018-02-05.

Werner Wolf. The Lyric: Problems of Definition and a Proposal for Reconceptualisation. In *Theory into Poetry. New Approaches to the Lyric*. Edited by Eva Müller-Zettelmann and Margarete Rubik. Pages 21-56. Rodopi ; Amsterdam. 2005. ISBN 90-420-1906-9.