Distinguishing Newspaper Genres. Exploring Automated Classification of Journalism’s Modes of Expression

Frank Harbers
f.harbers@rug.nl
University of Groningen, the Netherlands

Juliette Lonij
juliette.lonij@kb.nl
National Library of the Netherlands, the Netherlands

Introduction

This paper examines the opportunities, approaches and issues of automatically classifying historical newspaper articles from the Netherlands for ‘genre’ as an expression of the historically and culturally determined conception of journalism. Genre is defined as “language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms” (Handford 2010). As Barnhurst and Nerone (2001) argue: “The form includes the way the medium imagines itself to be and to act. In its physical arrangement, structure, and format, a newspaper reiterates an ideal for itself.” Examining the generic form of newspaper articles form a historical perspective therefore sheds an interesting light on the way newspaper journalism has developed.

The digital era, which is typified as the ‘age of abundance’ of historical newspaper material, poses new challenges to historical research. Historical approaches to selecting and analyzing newspapers, rooted in the assumption of a scarcity of available material, had to be replaced with social scientific methods, such as quantitative content analysis (Nicholson 2013; Broersma 2009). Yet, manual quantitative content analyses are still highly time consuming and therefore expensive. Moreover, even then the size of the material that can be covered represents only a small part of the amount of material available (Harbers 2014). Automated forms of (content) analysis have the potential to alleviate or even solve this issue. As such, automatic forms of content analysis would be highly suitable for longitudinal and also comparative historical research into the development of newspapers, which particularly grapples with the overabundance of available research material (Broersma 2009).

However, although these approaches have a great appeal to researchers (Allen, Waldstein & Zhu 2008; Grimmer & Stewart 2013), research in this vein is mostly done in information science and linguistics. It seldom has a press historical perspective (Broersma 2009). Moreover, the emphasis has mostly been on topical modeling (Lee & Myaeng 2002), whereas attention for automatic classification of style and genre is scarce. Rather than determining what subjects and themes are being discussed, this project aims to examine genre as a modes of expression of newspapers, shedding light on the discursive context (Handford 2010). This is a particularly difficult task as genres are dynamic and can change or fade away over time while new ones can emerge. Moreover, genres are ideal-typical discursive constructs, which means the textual manifestations do not always match the characteristics of these constructs perfectly, nor can they always be clearly delineated from other genres.

The research question therefore is:

To what extent and how can historical newspaper articles be automatically classified for genre?

To examine this question, we have designed a research project that builds on an existing set of metadata describing several textual characteristics, such as genre, of a large sample of historical newspaper articles. This dataset was the result of a large-scale research project into the historical development of European newspapers with the title ‘Reporting at the boundaries of the public sphere. Form, Style and Strategy of European Journalism, 1880-2005’. The set of metadata is derived from a manual content analysis that coded for genre based on a detailed rule-based coding manual. The metadata set relates to a corpus of approximately 33.000 Dutch newspaper articles from three types of Dutch newspapers, divided over the sample years 1885, 1905, 1925, 1965, 1985 and 2005 (Harbers, 2014). This set of metadata thus provides us with a number of labeled example articles that can be used to train and formally evaluate a classifier that is able to automatically predict the genre of additional samples of historical newspaper articles. In order to so the metadata **first** has to be connected to the full text of the corresponding digitized articles in the Dutch newspaper repository of the National Library (KB).

The **second** step is to use this enriched dataset to train a classifier that can classify historical newspaper articles automatically.

This paper first shows a way to connect the metadata to the digitized historical newspaper articles (**Phase 1**). Ultimately, it offers an outline of a concrete machine learning approach, applying linear and nonlinear classifiers, to predict the genre of a newspaper article. As a part of this, the paper discusses the different tools we have tried out and the problems we have encountered in the process (**Phase 2**). Specifically, the paper reflects on the way the rule-based approach to determining genre in the manual content analysis relates to the training of an automatic classifier based on machine learning techniques.

In order to link the articles described in the existing metadata set to the corresponding KB data, we first created a number of rules to find the most promising candidate links for each item in the original data set, based on the position of the article on the page, its size, and the presence of images and quotes. Since these rule alone turned out not to be sufficiently accurate, a simple classifier was trained to select the best link from the candidate set, if any, based on features such as the difference in size and number of images present between the article as described in the original data set and the article as found in the KB repository, as well as author mentions and subject matter, amongst other features. By only accepting links predicted by this classifier with a relatively high confidence value approximately 50% of all articles could be automatically linked, with an error rate of 0.5%. Some genres were underrepresented in the automatically linked set and were manually expanded upon.

The data set resulting from phase 1 was then used to create a training and test set for the actual genre classifier. This, again, involved taking several steps: first, the OCR for the articles in the set was retrieved and cleaned of quoted text by means of regular expressions. Next, the remaining text was pre-processed with the Natural Language Processing software package Frog, performing tasks such as segmentation, tokenization, and part-of-speech tagging. From the resulting annotated text, the relevant features for each article were calculated, such as number of words, number of sentences, number of direct quotes removed, and, most importantly, number of adjectives and various types of pronouns found in the text. These features, together with the existing genre labels were finally used to train a Support Vector Machine to choose one of eight possible genres for each article, ranging from news report and interview to news analysis and opinion article.

The initial results we obtained with this classifier were quite promising. Our first attempt resulted in an accuracy score of 58%, with the default accuracy by predicting the majority class or genre being 46%. The confusion matrix made from the predictions provides important information about which genres are most difficult to predict and what mistakes are most commonly made. This information can help us to fine-tune the existing features. Moreover, we have not yet implemented all the textual features that typify the different genres, such as named entities occurring in the text, sentiment, or self-classification to name only a few. Finally, we will compare the results of different algorithms, including deep learning approaches. Based on this, we expect to be able to improve these initial results in the coming months and present our final results in more detail at the conference.

Bibliography

Allen, R.B., Waldstein, I. & Zhu, W. (2008). 'Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres'. In: Buchanan, G., Masoodian, M.& Cunningham, S.J. (eds.), *Digital Libraries: Universal and Ubiquitous Access to Information*. New York: Springer

Barnhust, K. & Nerone, J. (2001). *The Form of News. A History*. New York: Guildford Press

Broersma, M. (2009). 'Nooit meer bladeren. Digitale krantenarchieven als bron'. In: *Tijdschrift voor Mediageschiedenis* 14(2): 29-55

Grimmer, J. & Stewart, B.M. (2013). 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. In: *Political Analysis* 21(3): 267-297.

Handford, M. (2010). 'What can a corpus tell us about specialist genres'. In: 'o Keeffe, A. & McCarthy, M. (eds.), *The Routledge Handbook for Corpus Linguistics*. New York: Routledge.

Harbers, F. (2014). *Between Personal Experience and Detached Information. The Development of Reporting and the Reportage in Great Britain, the Netherlands and France, 1880-2005*. PhD University of Groningen

Lee, Y. & Myaeng, S.H. (2002). 'Text Genre Classification with Genre-Revealing and Subject-Revealing Features'. SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 145-150.

Nicholson, B. (2013). ‘The Digital Turn’. In: *Media History* 19(1): 59-73.