Automating genre classification of historical newspaper articles. Mapping the development of journalism's modes of expression

This paper discusses a machine learning approach to automate the genre classification of Dutch historical newspaper articles and reflects on the challenges and its value. First, we discuss how we used an existing set of metadata to create a training set for the genre classifier and the challenges we faced in connecting the metadata to the original digitized historical newspaper articles. Subsequently, the paper outlines a machine learning approach to predict the genre of a newspaper articles, discussing and evaluating the different tools that were tested in the process.[1] Finally, it reflects on the way a traditional rule-based approach to determining genre relates to a machine learning approach.

Examining genre

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), genre can elucidate the underlying goals, norms and practices of journalism as a discourse. Examining journalistic genres from a historical perspective therefore elucidates how newspapers' conception of journalism developed. Yet, this type of longitudinal textual research is highly time consuming and still scarce. Moreover, the few attempts to systematically examine newspaper material, using social scientific methods such as quantitative content analysis, still only cover a fraction of the available material (Broersma, 2011; Harbers, 2014).

Automating such content analyses would be highly beneficial for research into the discursive development of newspaper journalism. This paper therefore critically discusses an approach to automate genre classification. This is a daunting task as genres are dynamic and can change or fade away over time while new ones can emerge. Moreover, genres are ideal types, which means the textual manifestations do not always match all the characteristics perfectly, nor can they always be clearly delineated from other genres.

A machine learning approach to automatically classify genre

Building on an existing set of manually coded metadata, describing several textual characteristics, such as genre, of a large sample (33.000) of historical newspaper articles[2], this paper outlines a machine learning approach to automate the genre classification of historical newspaper articles. This dataset thus provided us with metadata about a number of historical articles that was used to train and formally evaluate a classifier that is able to automatically predict the genre of additional samples of historical newspaper articles. Yet, the existing metadata needed to be linked to the corresponding digitized articles in the digital newspaper archives of the KB.

The paper will first discuss this linking process. We first selected the most promising candidate links for each item in the original data set, based on the position of the article on the

[1] The source code for training the classifier and applying it to new examples is available on GitHub (https://github.com/jlonij/genre-classifier) and everybody can experiment with the classifier through a graphical web interface created at http://www.kbresearch.nl/genre

[2] 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'.

page, its size, and the presence of images and quotes. A simple classifier was then trained to select the best link from the candidate set, if any, based on more precise features such as the size difference between the article and the candidate, as well as author mentions and subject matter. By only accepting links predicted with a relatively high confidence value approximately 50% of all articles could be automatically linked, with an error rate of 0.5%.

Subsequently, we will outline and discuss how the resulting data set was used to train the actual genre classifier. After the articles were pre-processed with the Natural Language Processing suite 'Frog', the annotated texts were examined for their textual features, including the length of the article, the number of direct quotes, the number of adjectives, various types of pronouns, and the number and position of named entities in the text. The selection of these features is based on the genre definitions of the codebook of the manual content analysis.

These features were used to train a classifier to choose one of eight possible genres for each article, ranging from news report to opinion article. We evaluated the performance through 10-fold cross-validation, using stratified sampling to create relevant subsets. A linear SVM classifier was chosen after comparison of various evaluation metrics with a number of other options (Naïve Bayes, non-linear SVMs and some simple neural networks), yielding the best results with an accuracy of 65%. It is important to note here that human coders do not always agree on what the right genre is. The intercoder agreement for genre in the manual content analysis was around 80% (Krippendorf's alpha, taking into account chance, was between 0.7 and 0.8 in different groups of coders). As such, 65% is considered a very promising result.

Finally, we reflect on the relation between a rule-based and machine learning approach to the classification of genre. We will discuss the significance of individual features in the machine learning process and show how the 'confusion matrix' provides valuable information about the common mistakes of the classifier and which genres are most difficult to predict. Moreover, as the probability for the predicted genre as well as for the other genres is known, we will discuss how these numbers offer insights in the dynamic nature of journalistic genres.

Bibliography

- Broersma, M. (2011). 'Nooit meer bladeren. Digitale krantenarchieven als bron'. In: *Tijdschrift voor Mediageschiedenis* 14(2): 29-55

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