Whose 'I' is it anyway? Comparing a rule-based approach and a BERT token-classifier for quote detection in Dutch newspapers

Long paper submission, intending to present in person

Kim Smeenk[1], Herbert Kruitbosch[2], Frank Harbers[1] & Marcel Broersma[1]
[1] Centre for Media and Journalism Studies, University of Groningen
[2] Centre for Information Technology, University of Groningen

1. Introduction

With journalism's current struggle to maintain its authority and commercial viability, alternative norms and practices have become more central in debates on journalism. Embracing a stronger personal engagement in covering the news has become a prominent strategy to (re)engage audiences and foster their trust. Often subjectivity is incorporated into the story by quoting sources that share their personal experiences (Wahl-Jorgensen, 2013). But personal journalism, in which journalists explicitly foreground their own subjectivity, has also been increasingly present in Dutch newspapers. Personal journalism deviates from what is commonly understood to be trustworthy journalism, which is supposed to be neutral and objective. The increase of personal journalism therefore could have far-reaching consequences, as it potentially changes what is considered to be trustworthy journalistic knowledge (Tulloch, 2014). Because of the abundance of newspaper data and the time consuming nature of close reading, we apply computational methods to find personal journalism in newspapers.

We operationalize personal journalism as all journalism in which the journalist explicitly refers to themselves using first-person pronouns. To successfully extract personal journalism articles from newspaper articles, we need to be able to distinguish between first-person pronouns that refer to the journalist and first-person pronouns that refer to their sources. While it is relatively easy to filter opinion articles and letters to the editors from the dataset based on the available metadata, the first-person pronouns in quotes need to be distinguished by automatically extracting them from the text.

In this paper we present a comparison of computational approaches to identifying and extracting quotations from newspaper articles – a necessary first step in the attempt to automated detection of personal journalism. We compare a rule-based system and a machine learning system using BERT.

2. Related work

Quote extraction has previously been approached from two perspectives: rule-based and machine learning.

The rule-based approach uses regular expressions, hand-crafted rules or hand-crafted linguistic features to extract quotes from texts (Pouliquen et al., 2007.; Särg et al., 2021; Sarmento & Nunes, 2009; van Atteveldt, 2013). For direct quotes, this normally relies heavily on finding quotation marks. This approach has as its major advantage that the performance is usually very high (precision of above 0.90). The drawback can be that the performance is hampered for texts that do not consistently use quotation marks or other linguistic features. Moreover, while the precision of these systems is usually very high, the recall is lower. While this is not so problematic for studies that attempt to extract quotes to study the content of those quotes, for us it is more important to have high recall on the quotes, as it will aid us in maximising precision in detecting personal journalism.

Machine learning approaches to quote extraction use supervised or unsupervised models to automatically detect direct speech (Byszuk et al., 2020.; Pavllo et al., 2018; Purnomo et al., 2021). While the results of these models are usually not as good as rule-based systems, recent experiments with deep learning models have shown promising results on 19th century novels (Bysuk et al. 2020) and have been suggested as a fruitful framework for newspaper texts (Purnomo et al. 2021).

To see which approach is the most fruitful for our purpose, we compare this rule-based system and deep learning approach on Dutch newspaper articles. This comparison makes a valuable contribution to the fields of digital methods and computational journalism studies as this is to our knowledge the first study to apply a deep learning model to newspaper data for quote extraction. Moreover, we do not know a study that compares these two methods for journalistic texts.

3. Data

We have collected a sample of two constructed weeks (Riffe, Aust & Lacy, 1993) from three Dutch newspapers (*Algemeen Dagblad*, *NRC Handelsblad* and *Volkskrant*) for 1999, 2009 and 2019 using the Nexis database. This dataset contains 12760 articles. Two annotators have manually labelled each article for the binary label 'personal journalism', deciding whether at any point in the article the journalist refers to themselves as 'I' or 'we'. The annotators were provided with clear predefined guidelines to decide when 'I' and 'we' refer directly to the journalist. In total 958 articles were labelled as personal journalism. A subset of 250 articles has been annotated with in-text annotations, for the labels 'I', 'we' and 'quote'. The first two were annotated for all first-person pronouns that referred directly to the journalist. The latter for all direct quotes in the text. We have randomly selected 200 articles of personal journalism and 50 articles of other journalism, to ensure that we had enough labelled data for the 'I' and 'we' labels.

	# tokens	# I	# we	# quote	quote ratio
Personal journalism	96093	712	203	496	0.115
Other journalism	15551	0	0	101	0.127

Table 1: Corpus characteristics

4. Method

We have designed a rule-based and a deep learning method for quote extraction, and performed two experiments with both these methods based on two ways of text segmentation. As we observed that quotation marks were often omitted if the end of the quote coincided with the end of the paragraph, we applied both methods to the article texts and to separate paragraphs.

For our rule-based system, we use a regular expression that takes into account the different quotation marks and combinations of them that are used: "…"; '…', ‚…', ‚…", "…', '…". We have applied our regular expression twice, to maximise our chance of capturing quote-in-quotes.

The manually labelled data was used to train a classifier that is able to recognize quotes and journalists personal pronouns in unseen news articles. Like Byszuk et al. (2020) we have adjusted a token-classification method using BERT to detect quotes. BERT is a pre-trained language representation model that can be fine-tuned for specific tasks, such as token classification (Devlin et al. 2018). The classification task consists of labelling tokens with one of the following labels: 'ik', 'we', 'quote' while also allowing to detect the start and end of a sequence of 'ik', 'we' or 'quote' tokens. As Byszuk et al. (2020) noted that first-person narration often gets wrongly labelled as direct speech, we have added the first-person pronoun labels to signal pronouns that are outside of quotes.

5. Results

5.1 Performance

We validate the results of our approaches in two ways: by comparing the performance with our labelled data and by evaluating which is more effective for the task that we are eventually interested in: detecting personal journalism. We are currently working on the evaluation of the methods on both metrics. Here we report the results of the BERT token-classifier on the first task and the results of the rule-based method on the second task. These are not yet comparable, but give a first impression of

their performance. These first results (table 2 and 3) suggest that the BERT token-classifier is more effective. The rule-based method does manage to correctly detect personal journalism, but only 41% of the journalism that is detected as personal journalism truly falls into that category. Further evaluation will point out definitive conclusions.

	Per article	Per paragraph
Precision	0.85	0.99
Recall	0.80	0.98

Table 2: evaluation of BERT token-classifier on quote extraction and first-person pronoun detection

	Personal journalism	Other journalism
Precision	0.41	0.99
Recall	0.99	0.91

Table 3: evaluation of rule-based system on personal journalism classification

5.2 Error analysis

Our error analysis of the BERT token-classifier suggests that three errors are the most frequent.

1. **Sliding window errors**
   The majority of the errors are correctly recognized quotes whose spans have not been detected correctly, as the start or the end of the quote are left out, or extra words are added. We suspect this is a result of the way BERT truncates the data and works with a set amount of tokens at the same time.
2. **Conflicting labels**
   There are some examples of quotes that have been correctly detected and only the first-person pronouns within the quote are detected as 'I' or 'we' labels. This happens especially when there are many quotes in the text. This suggests that our attempt to minimise first-person narration errors has been counterproductive.
3. **Unclear**
   A few words are seemingly randomly assigned the 'quote' label. A pattern is hard to distinguish.

Our error analysis of the rule-based method suggests that, as expected, it is mostly sensitive to inconsistent use of quotation marks.

1. **Quote-in-quote**
   After running twice, the rule-based method was better at capturing quote-in-quotes, but these were still often only partly recognized. Moreover, by running it twice, other errors were introduced, such as partial recognition of quotes, or long additions to the end of the quote span.
2. **No opening or closing quotation mark**
   If there were inconsistencies in the text, and quotation marks were missing at either end of the quote this resulted in either not detecting the quotation, or wrongly detecting long strings of text as quotations.

6. Conclusion

In this paper we have presented two solutions to quote extraction from Dutch newspaper texts with the purpose to investigate the possibility of automatically extracting personal journalism. Our first results suggest that our rule-based system does not reach high enough precision on our eventual goal to detect personal journalism. For now, our BERT token-classifier shows promising results in quote extraction that suggest that it would perform well in detecting personal journalism.

For future work we aim to train the classifier without the personal pronouns labels and try out different sliding windows to minimise the error rates.

Bibliography

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., & Eder, M. (2020). Detecting direct speech in multilingual collection of 19th-century novels. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 100-104).
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Pavllo, D., Piccardi, T., & West, R. (2018, June). Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping. In *Twelfth International AAAI Conference on Web and Social Media*.
Pouliquen, B., Steinberger, R., & Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 487-492).
Purnomo W.P., Y. S., Kumar, Y. J., & Zulkarnain, N. Z. (2021). Understanding quotation extraction and attribution: Towards automatic extraction of public figure's statements for journalism in Indonesia. *Global Knowledge, Memory and Communication*, *70*(6/7), 655–671. https://doi.org/10.1108/GKMC-07-2020-0098
Riffe, D., Aust, C. F., & Lacy, S. R. (1993). The effectiveness of random, consecutive day and constructed week sampling in newspaper content analysis. *Journalism quarterly*, *70*(1), 133-139.
Särg, D., Kink, K., & Masing, K. O. (2021). Quote extraction from Estonian media: Analysis and tools. *Eesti Rakenduslingvistika Ühingu aastaraamat*, *17*, 249-265.
Sarmento, L., & Nunes, S. (2009). Automatic extraction of quotes and topics from news feeds. In *DSIE'09-4th Doctoral Symposium on Informatics Engineering*.
Tulloch, J. (2014). Ethics, trust and the first person in the narration of long-form journalism. *Journalism: Theory, Practice & Criticism*, *15*(5), 629–638. https://doi.org/10.1177/1464884914523233
Van Atteveldt, W. (2013). Quotes as data extracting political statements from Dutch newspapers by applying transformation rules to syntax graphs. In *Text as Data Conf. London, 2013* (pp. 1-9).
Wahl-Jorgensen, K. (2013). Subjectivity and story-telling in journalism: Examining expressions of affect, judgement and appreciation in Pulitzer Prize-winning stories. *Journalism Studies*, *14*(3), 305-320.