Niels Bijl

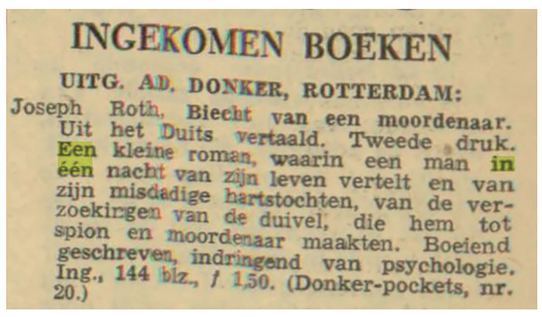
BOOK TITLE EXTRACTION
A Named Entity Recognition Approach

INTRODUCTION

Background: Newspaper book reviews provide insights into cultural and intellectual trends. Accurate extraction of book titles from these reviews can enhance cultural and literary analysis.
Current Challenges: Existing rule-based methods are limited by accuracy and generalizability, necessitating extensive manual verification.
Objective: Develop an autonomous system utilizing Named Entity Recognition (NER) to extract book titles from OCR-scanned historical newspapers.
Research Question: *To what extent can Named Entity Recognition be utilized to autonomously extract book titles from OCR-scanned historical newspapers, thereby facilitating deeper cultural and literary analyses?*

DATA

Leeuwarder Courant: 12,535 book reviews from 1962 till 1995, containing 23,529 book titles.
Trouw: 115 book reviews.
Het Parool: 193 book reviews.
Challenges:
- Vaulty Optical Character Recognition Quality.
- Leeuwarder Courant data not labeled for NER.


INGEKOMEN BOEKEN

UITG. AD. DONKER, ROTTERDAM:

Joseph Roth. Biecht van een moordenaar. Uit het Duits vertaald. Tweede druk. Een kleine roman, waarin een man in één nacht van zijn leven vertelt en van zijn misdadige hartstochten, van de verzoekingen van de duivel, die hem tot spion en moordenaar maakten. Boeiend geschreven, indringend van psychologie. Ing., 144 blz., f. 1,50. (Donker-pockets, nr. 20.)

INGEKOMEN BOEKEN
UITG. AD. DONKER, ROTTERDAM-
JOS6?_It h^ thV_^ eeit Tan 'en moordenaar.
elÏ k^ln^ ltS T«*aal°. Tweede d"k.
-ïn ,!__?_2 e roman' "aarin ««" man in
Sn SL™. ? Jn teve" vertelt eTvan
foëkiS^^ hartstochten, van de ver-
sokrn^n m^ .f* duivel- di« hem tot
g^^en^ri^^3^"- Boelend
Ing M^ 7 m? r 1 m«<m<k v? n Psychologie,
ing., I*4 wz„ / i oo (Donker-pockets. nr.

Figure 1: Example of inaccurate OCR from the Leeuwarder Courant (30-06-1958) showing the original text (left) and the erroneous OCR output (right).

METHODOLOGY

NER Models: Evaluated multiple NER models:
- **SpaCy Baseline:** Pre-trained model for extracting the entity WORK_OF_ART.
- **Training SpaCy:** Fine-tuned pre-trained model on our dataset.
- **BiLSTM-CRF:** Utilized FastText embeddings with hyperparameter tuning for memory units (50, 100, and 200).
- **Transformer-based Models:** Fine-tuned several pre-trained Dutch NER transformers from HuggingFace (e.g., BERT, RoBERTa).

Evaluation
- **Metrics:** F1 score, precision, and recall.
- Token classification.
- Matching to the Nederlandse Bibliografie Totaal (NBT).

REFERENCES:
Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., & Doucet, A. (2019, June). An analysis of the performance of named entity recognition over OCRed documents. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 333-334). IEEE.

RESULTS

BiLSTM-CRF: Best performance with 200 memory units.
Token Classification: Transformer model "xlm-roberta-large-finetuned-conll03-english" achieved the highest F1 scores: 83.9% on validation data, 84.3% on the LC test set, and 56% on the Trouw & Parool test set.
Book Title Extraction: Matching NER output to the NBT yielded an F1 score of 59.4% on the LC data and 57.2% on the Trouw & Parool data.

Named Entity Recognition model	Leeuwarder Courant validation			Leeuwarder Courant test			Het Parool & Trouw test		
	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)
Baseline (SpaCy)	9,6	12,8	7,7	9,1	12,5	7,2	21,1	31,6	15,9
Trained Spacy	64,9	74,7	57,3	63,9	74,3	56	32	71,8	20,6
BiLSTM-CRF (200 memory units)	68,8	73,8	64,4	69	74,8	64	34,6	77,2	22,3
xlm-roberta -large-finetuned -conll03-english	83,9	82,9	85	84,3	83,4	85,2	56	78,7	43,3

Figure 2: Final token classification results: F1 score, precision, and recall on the Leeuwarder Courant validation set, Leeuwarder Courant test set, and Het Parool & Trouw test set.

Dataset	F1 (%)	Recall (%)	Precision (%)
Leeuwarder Courant test	59,4	54,1	65,9
Het Parool & Trouw test	57,2	50,3	66,4

Figure 3: Performance of the xlm-roberta-large-finetuned-conll03-english model in matching book titles from the NBT: F1 score, recall, and precision on the Leeuwarder Courant and Het Parool & Trouw test datasets.

DISCUSSION

OCR Impact: Faulty OCR likely affected NER performance (Hamdi et al., 2019), though the extent remains unknown.
Model Strengths & Weaknesses: High precision indicates few false positives, but lower recall suggests missed book title tokens, often due to incomplete book titles in training data.
Dataset Discrepancies: The Leeuwarder Courant data is not annotated directly in a NER format, while the Parool & Trouw dataset is, with multiple annotations for the same title in the latter leading to a lower recall score as the model predicts only a single instance of each title.
Disappointing Matching to NBT: Despite high NER performance, matching results to the NBT were suboptimal. Accurate matching requires more than the main title alone.

CONCLUSION

Transformer-based large language model can accurately and autonomously extract text representing book titles from book reviews within historical newspapers. However, accurate matching to the NBT requires more than just main title.
Future Work: Annotating datasets directly for NER and exploring nested NER for including additional details.