

Event Extraction from Radio News Bulletins for DIVE+

Kim van Putten - Vrije Universiteit Amsterdam
Supervisor: Oana Inel

Problem Definition

DIVE+ is an event-centric linked data digital collection browser, but in order to interlink the data we need to identify the events first.

Event extraction from radio bulletins has proven to be problematic. The text content of the bulletins extracted by OCR contains errors, Named Entities extracted from the text are not always correct, and the extracted events in the metadata are a best guess.

Research Question

Can we find a better way to extract events from the bulletins to improve the linkage?


The Dataset

215 radio news bulletins with metadata consisting of:

- Text content
- Date
- Named entities
 - Actors - 1296 entities
 - Places - 526 entities
 - Other related concepts - 526 entities

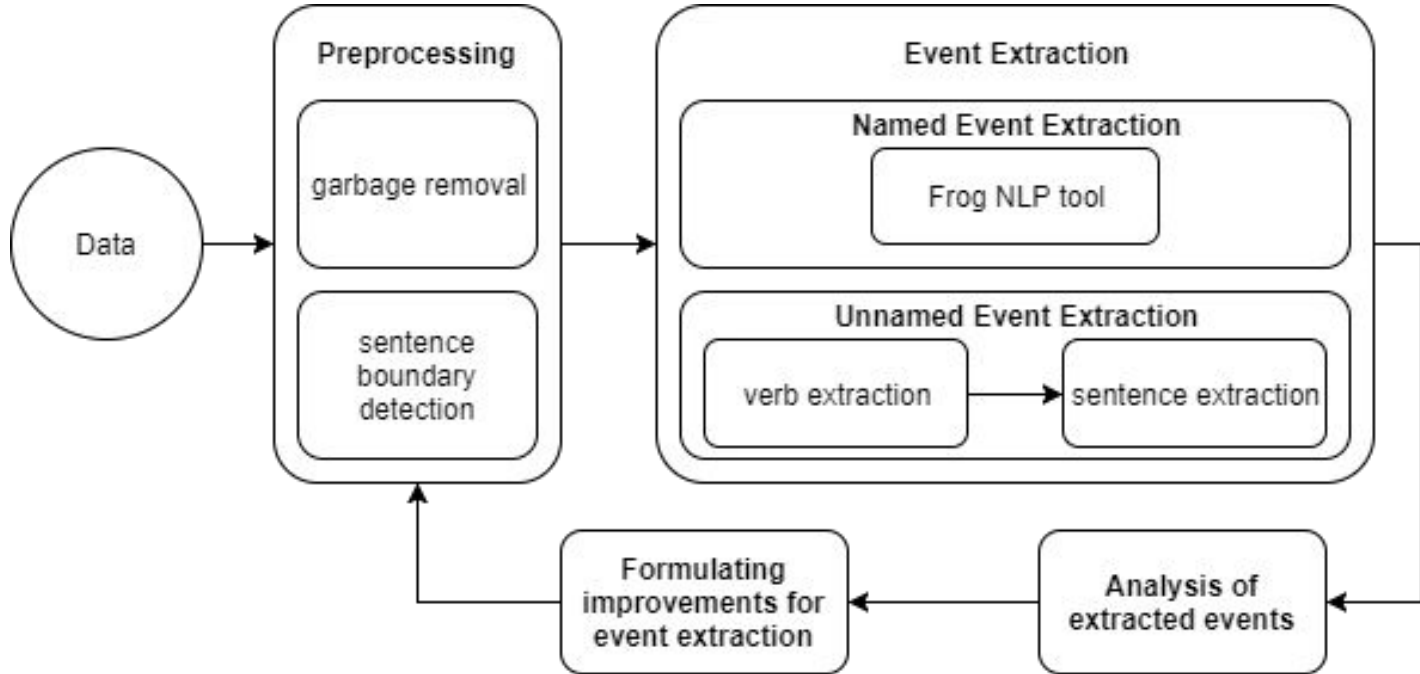
E.g. *HoHand-Amerika-lijn, Barakkenkamp, Foreign Office.*

- Event - 215 entities



Methodology

- Extract named events with NER tool Frog.
E.g. *Tweede Wereldoorlog, Olympische Spelen, demonstratie*
- Extract unnamed events by finding the first sentence that contains a verb, actor and/or place:
[Someone doing something somewhere](#)
E.g. *“Een aantal hoge functionarissen uit Spaansch Marokko is in RABAT aangekomen”*
- Iteratively formulate, implement and analyse improvements to achieve the best possible event extraction.



Results

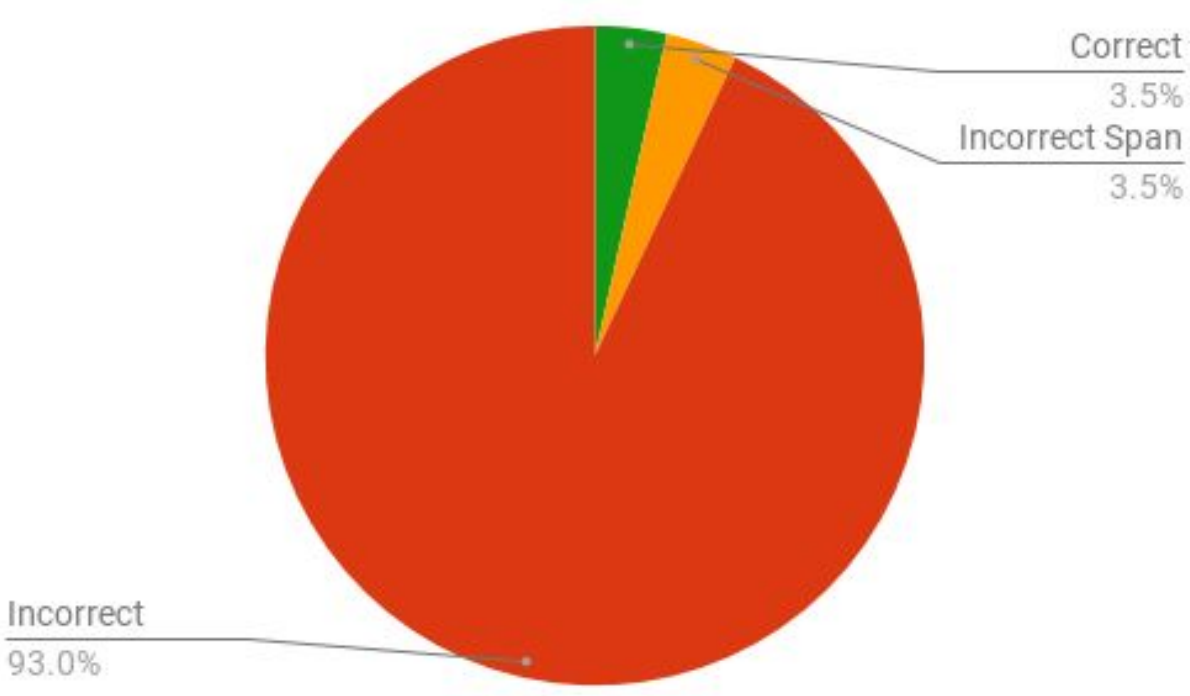


Figure 1: Evaluation of named events extracted with Frog

Frog appears to perform poorly on the OCR text. Only 57 named events were extracted, and only 3.5% were correct.

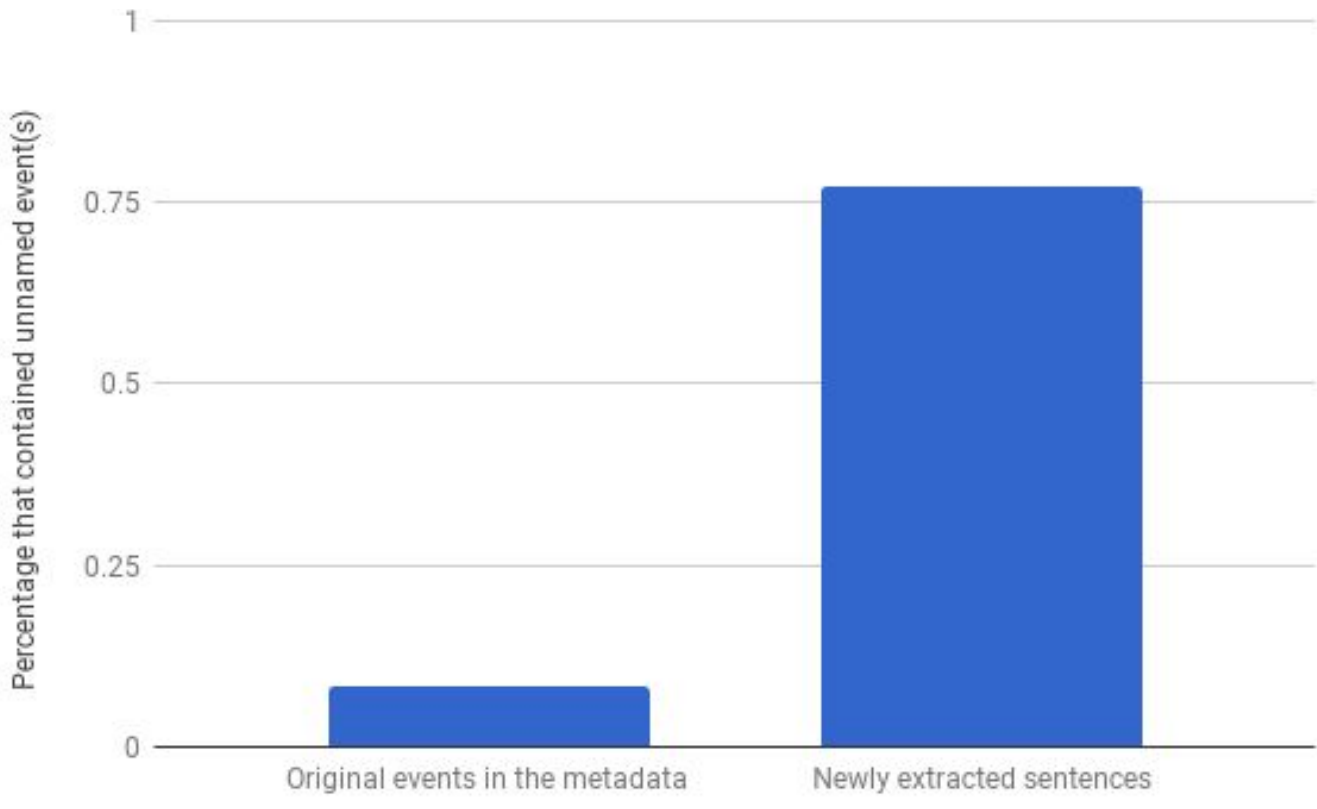


Figure 2: A comparison between the old events in the metadata and the new unnamed events extracted with the new method. The figure compares the percentage of the sentences that were eventful.

By searching for verbs, actors and places, we extract more sentences that contain a historic unnamed event (77.2%) compared to the original events in the metadata (8.4%).

Conclusion

- By searching for verbs, actors and places we improved upon the old event extraction method.
- However, extracting sentences is very *coarse-grained*.
 - The machine cannot tell how many events there are in an extracted sentence.
 - We cannot automatically detect which verbs are related to which entities.
- The event extraction method only extracts one event per bulletin.
- The event extraction method *did not improve linkage* in the data structure, but *did improve searchability* of media objects because the new events contained more likely search terms.
- Errors in OCR are a big obstacle: sentence boundary detection, NER and verb extraction all suffer in performance.

A more fine-grained event extraction method is needed, but a fully pattern-based approach is not sufficient to for this task.

