Event Extraction From Radio News Bulletins

Kim van Putten
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
ke.vanputten@gmail.com

Oana Inel
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
oana.inel@vu.nl

Victor de Boer
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
v.de.boer@vu.nl

Lora Aroyo
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
lora.aroyo@vu.nl

1 INTRODUCTION

Exploratory search systems provide information to users with an unclear information need, by offering support for browsing strategies through carefully designed interfaces that support interactive forms of search [3]. DIVE+[1] is a linked data digital cultural heritage collection browser that organizes historical media objects and facilitates exploratory search through event-centric linking of the data [2]. The DIVE+ browser facilitates exploration and learning through an intuitive and interactive interface which allows the end user to browse media objects from four heritage institutions (Netherlands Institute for Sounds and Vision[2], Dutch National Library (KB)[3], Amsterdam Museum[4] and Tropenmuseum[5]). All objects have metadata which includes descriptive text, related entities such as actors, places and events and links between them.

In the DIVE+ project, event extraction proved to be particularly difficult for KB media objects, *i.e.*, radio news bulletins (see Figure 1. On the one hand, the media objects in the KB dataset deal with several issues (see footnote 6) introduced by the OCR software: *(1)* garbage strings, *(2)* misidentified characters and spelling errors. Due to these errors, Natural Language Processing (NLP) tools also struggle to extract meaningful entities [1]. On the other hand, the NEs in the metadata are not always correct or identified (*e.g.*, *Staatsoourant* was incorrectly classified as actor, the place *DEN HAAG* was not extracted). Considering these and the fact that currently, the KB objects do not have a well defined event, we formulate the following research question: *"Can we find a more effective way to extract events from the KB radio news bulletins to improve linkage within the DIVE+ demonstrator?"*. We aim to find a better approach to extract events from the KB dataset rather than extracting the first 100 characters.

[1] http://diveplus.frontwise.com/
[2] http://www.beeldengeluid.nl/
[3] https://www.kb.nl/
[4] https://www.amsterdammuseum.nl/
[5] http://www.opencultuurdata.nl/wiki/tropenmuseum/

SEMANTiCS 2017 workshop proceedings: EVENTS
September 11-14, 2017, Amsterdam, Netherlands

2 METHODOLOGY

This section describes our research methodology for finding a more suitable method to extract events from the radio bulletins. We apply our research methodology on a subset of 215 news radio bulletins from KB dating from April 1939.

2.1 Preprocessing

As mentioned previously, the content of the bulletins contains errors caused by OCR. Therefore, we first perform garbage removal from the text of the bulletins by adopting a series of pattern-based approaches from [4]. When a string is identified as a garbage string, it is removed from the text. Second, we perform sentence boundary detection by assuming that all sentences end with a period.

2.2 Event Extraction

We distinguish two types of events: *named events* and *unnamed events*. Named events are events which have a name, *e.g.*, *Olymische Spelen*. Unnamed events are linguistic events, which do not have a name, *e.g.*, the sentence *"functionarissen uit Spaansch Marokko is in RABAT aangekomen."* describes the event of arriving in Rabat.

Named Event Extraction: To extract named events from the bulletins, we used the NLP system Frog[6]. When Frog recognizes a token in the text as a NE, it assigns it a type (*i.e.*, person, organization, location, product, event or miscellaneous). To identify the events, we extracted only the tokens which have been typed by Frog as events.

Unnamed Event Extraction: Since unnamed events can not be detected in texts with typical NER tools, we first identify actions by means of verbs, using the NLP tool TreeTagger[7]. We are interesting in identifying *eventful sentences*, *i.e.*, sentences that contain one or more unnamed events under the pattern *someone, doing something, somewhere*. We attempt to extract unnamed events from the bulletins using a knowledge-driven approach which exploits the NEs already in the metadata of the bulletin and the actions (*i.e.*, verbs) identified by TreeTagger. Since not all events might

[6] https://languagemachines.github.io/frog/
[7] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

be associated to both an actor and a place, we introduce a tiered method of sentence extraction: (1) Tier 1: Sentence contains at least a verb, an actor and a place; (2) Tier 2: Sentence contains at least a verb, and either at least an actor or a place and (3) Tier 3: Sentence contains at least a verb.A sentence matching tier 1 is more likely to contain an unnamed event than a sentence of tier 2, and thus, tier 1 is preferred over tier 2 and tier 2 is preferred over tier 3. If there are no sentences that have at least one verb, then we keep the initial string of the first 100 characters as the event.

3 RESULTS

In this section we briefly present the results of all the intermediate steps of the event extraction pipeline.

3.1 Preprocessing Results

Overall, 2,574 garbage strings were removed from the data. Despite the large number of strings removed, there are still garbage strings undetected. All attempts of adding new rules or changing the current ones to recognize similar words resulted in the seesaw phenomena, where the removal of garbage strings also led to the removal of non-garbage strings.

3.2 Event Extraction Analysis

The extracted named events and eventful sentences are analyzed to determine how well the extraction methods perform.

Analysis of Named Event Extraction: Frog extracted a total of 57 events from the 215 bulletins in the data. Overall, it appears that Frog performed very poor on event extraction. Only 4 out of 57 extracted events are actually events, and 2 out of these 4 have an incorrect span.

Analysis of Unnamed Event Extraction: We extracted one sentence per bulletin using the 3-tier extraction method: 92 sentences in tier 1, 85 sentences in tier 2, 15 sentences in tier 3 and for 23 bulletins we found no sentence containing a verb so we kept the initial event. Further, we manually evaluate the sentences extracted with the 3-tier method and compare them with the baseline, *i.e.*, the current event strings in the metadata. A string was considered an event if *(1)* it was reported as something that happened, is happening, or will/may happen at a later date, *(2)* it is based on a verb or a set of verbs, and *(3)* it has historic value.

We found out that from the original events which were extracted by taking the first 100 characters only 8.4% contained unnamed events. From the sentences that were extracted with the new 3-tier method 77.2% were eventful. Thus, the new method of extraction provides better suited events mentioned in the bulletins. The event strings in the metadata that were found eventful contained exactly one event. The newly extracted eventful sentences contained more than one event on average (1.5 events) which means that overall, they are more expressive. We identified the following reasons why a sentence did not contain an unnamed event: *(1)* no new sentence was extracted because TreeTagger did not recognize any verbs in the text of the bulletins (bulletins without verbs or misspelled verbs); *(2)* words were incorrectly tagged as verbs; *(3)* incorrect sentence boundary detection and *(4)* incorrect NE in the metadata. For the first two observations the quality of the OCR negatively impacts the performance of the event extraction. We address the fourth observation in the next part.

3.3 Improvements of the Event Extraction

The extraction of unnamed events relies on finding relationships between verbs and NEs. However, in Section 1 we see that the NEs in the metadata of the bulletins are not always correct. Overall, we see that about a quarter of all the NEs are incorrect or mistyped. Actors show to have the largest percentage of correct NEs but simultaneously the largest percentage of incorrect extracted NEs (15.4%). Next, we investigated whether we can improve the named entities from the bulletins using Frog. Frog extracted a total of 5,807 NEs of type person, organization, location and event. However, we see that Frog performs poor on the extraction of all entity types (only around 20% were correct) and hence, we chose not to use the NEs extracted by Frog in our pipeline.

We further analyzed two assumptions: *(1) Sentences that contain a verb, an actor and a place are more likely to contain unnamed events than sentences which do not have both an actor and place.* To prove this assumption we tested the unnamed event extraction with a 2-tier method which is identical to the 3-tier method except that we omit tier 1. We evaluate the sentences extracted by the 2-tier system and compare them to the sentences extracted by the 3-tier method. Results show that fewer of the sentences extracted by the 2-tier method are eventful compared to the 3-tier method (0.65 compared to 0.8). *(2) The main event or most important events are mentioned at the beginning of the text.* We conclude that limiting the extraction to a specific part of the text results in slightly worse event extraction because *(1)* the extractor might be forced to extract a sentence that matches a lower tier and *(2)* if a text contains only one sentence with a verb an actor and a place, we may not chose the part of the text where this sentence is placed.

4 CONCLUSION

This paper presents a methodology to extract events from radio news bulletins to improve the exploratory search offered by DIVE+ using a NER tool and a pattern-based approach which exploits the NE space in the metadata of the bulletins. The new events are full sentences, less likely to be header information of the bulletin and more likely to include relevant NEs and terms that a user might search for. Therefore, the bulletins are more likely to show up in search results (see Figure 2). On the one hand, the NER tool Frog was unsuccessful at extracting events from the radio bulletins. On the other hand, the pattern-based method improved the events, which was further beneficial for the searchability and the presentation of the media objects. Overall, errors in the OCRed data turned out to be problematic for sentence boundary detection, NE extraction and ultimately for the extraction of events. To achieve a finer granularity of event extraction, future research is necessary to identify what is the relationship between the NEs and the verbs that describe an event. It might also be worthwhile to invest further research in OCR post-correction and normalization to improve the quality of the data so that better NER can be achieved.

REFERENCES

[1] Beatrice Alex and John Burns. 2014. Estimating and rating the quality of optically character recognised text. In *DATeCH*. ACM, 97–102.

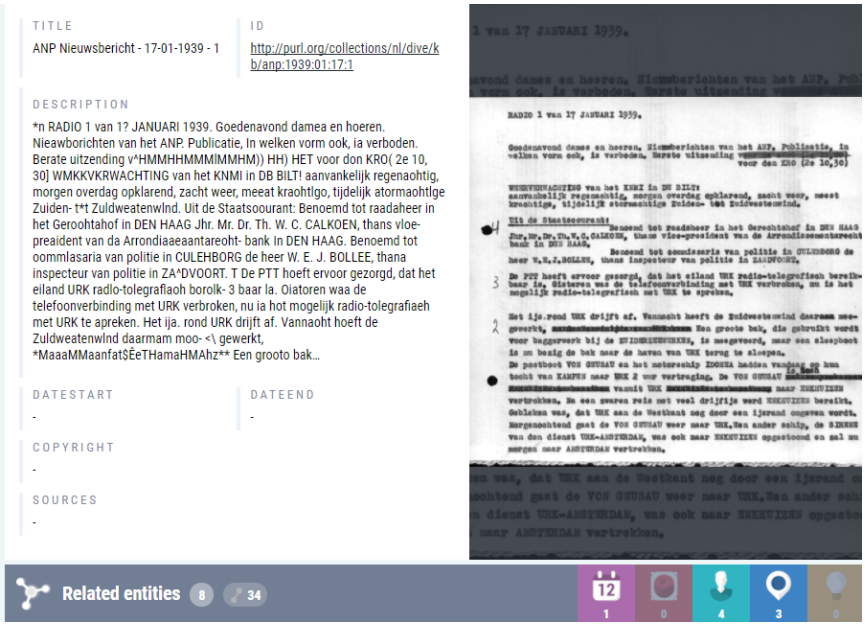

Figure 1: An example of an ANP radio news bulletin in the DIVE+ demonstrator

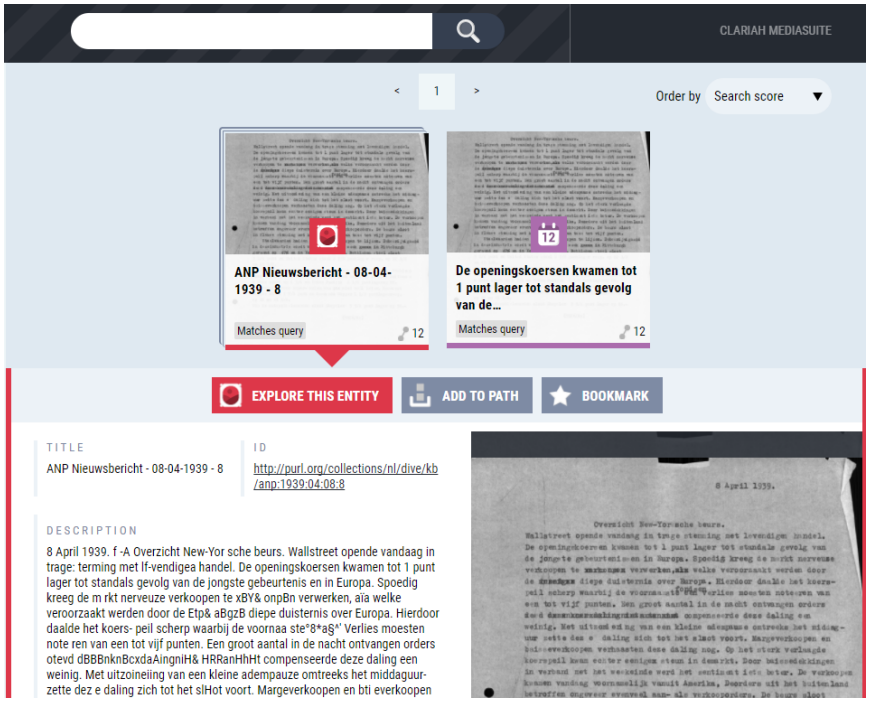

Figure 2: The search result for the query "openingskoersen 1 punt lager" in the DIVE+ demonstrator after the data enrichment with the new events. The left object shows a radio bulletin and the right object is the event associated with the bulletin.

[2] Victor De Boer, Johan Oomen, et al. 2015. DIVE into the event-based browsing of linked historical media. *Web Semantics: Science, Services and Agents on WWW* 35 (2015), 152–158.

[3] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[4] Kazem Taghva, Tom Nartker, Allen Condit, et al. 2001. Automatic removal of "garbage strings" in OCR text: An implementation. In *WMSCI*.