

Hacking History via Event Extraction

Roxane Segers
Department of Computer
Science
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The
Netherlands
r.h.segers@vu.nl

Marieke van Erp
Department of Computer
Science
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The
Netherlands
marieke@cs.vu.nl

Lourens van der Meij
Department of Computer
Science
VU University Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The
Netherlands
lourens@cs.vu.nl

ABSTRACT

Within cultural heritage collections, objects are often grounded in a particular historical setting. This setting can currently not be made explicit, as structured descriptions of events are either missing or not marked up explicitly. This paper reports a study on automatic extraction of an historical event thesaurus from unstructured texts. We show how this preliminary thesaurus accommodates event- and object-driven search and browsing of two cultural heritage collections.

Categories and Subject Descriptors

I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods; I.2.7 [Artificial Intelligence]: Natural Language Processing; J.5 [Computer Applications]: Arts and Humanities

General Terms

information extraction, event modeling, cultural heritage

1. INTRODUCTION

Events have recently gained attention in the knowledge representation community as valuable constructs [4, 7, 8] that can help tie together relevant but yet unrelated elements of information. In the cultural heritage domain, knowledge about historical events is often concealed in textual descriptions that can only be accessed via keyword search. As such, the available knowledge can not be reused across collections as it is not part of the shared metadata and controlled vocabularies.

In this study, we investigate how historical events in unstructured text collections can be captured and modeled to create an event thesaurus for enriching metadata in cultural heritage collections. We adopt the SEM event model [8] to distinguish event types, actors, locations, and dates. We experiment with natural language processing (NLP) techniques to extract event names and their associated actors, dates and locations. Additionally, we show how this resulting preliminary event thesaurus is employed in a new platform for event- and object driven searching and browsing of the collections of the Rijksmuseum Amsterdam (RMA) and the Netherlands Institute for Sound and Vision (S&V).

2. EVENT EXTRACTION FROM TEXT

As no annotated historical document collections exist in Dutch, our approach is focused on extracting named events with minimal manual effort. For this study we selected 3,724 historical Wikipedia articles as a test set. The event extraction process consists of three steps: in the **first step**, we recognize *actor names* and *locations* using the Stanford Named Entity Recognition system [2] adapted for Dutch historical texts. Dates were recognized via regular expressions. This step resulted in 18,623 candidates for actors (F-measure of 0.77), 7,023 locations (F-measure of 0.66) and 7,981 dates. In the **second step**, we use a pattern-based method for recognizing *event names* such as *French Revolution*. We harvest patterns from the Web (e.g., *destroyed during the, before the*) using the Yahoo! search API ¹ and a seed of one hundred historical events. Patterns are ranked by frequency of co-occurrence with two or more seed events [6]. To retrieve event candidates, we applied the patterns to the Wikipedia corpus. The event candidates are then filtered, based on a threshold on the pattern score, resulting in a set of 2,444 unique events. The precision score of this set is 56.3%.

In the **third step**, we associate events with actors, locations and dates. We experiment with both redundancy and co-occurrence of data on the Web, inspired by the work of Geleijnse et al. [3] and Cilibrasi & Vitanyi[1]. Each combination of an event name and actor/location/date is sent to Yahoo! and for each pair a score is computed. We discovered 392 event names that were paired with an actor, a location and a date. Through manual evaluation we conclude the following: 71.9% (323) are correct event names, 45.6% (179) are correct actors, 41.1% (161) are correct locations and 51.5% (202) are correct dates.

3. ENRICHMENT BY EVENTS

The extracted events are linked to the RMA and S&V collections. In total 35 unique events provide direct relations from 435 S&V objects to 675 RMA objects. An additional 34 unique events provide links from 391 S&V objects to 362 RMA objects, but this link exists indirectly through the event instance (e.g., S&V object - Actor - RMA object). We hypothesize that these links are potentially useful for navigating cultural heritage collections.

Copyright is held by the author/owner(s).
K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.

¹<http://developer.yahoo.com/search>

Slachtoffers gemaakt door de Nederlandse troepen op weg naar Jogjakarta¹ (Object)



Slachtoffers gemaakt door de Nederlandse troepen op weg naar Jogjakarta. Kinderschilderij van de inname van Jogjakarta tijdens de tweede politionele actie, december 1948.
NG-1998-7-10

Associated Events
DepictsEvent: [Tweede politionele actie¹](#)

biographical aspects
Creator: Toha Adimidjojo, Mohammed¹(4) Date: 1948-12-19¹(3) 1949-06-30¹(3) 20e eeuw¹(18) tweede kwart 20e eeuw¹(17)

material aspects	semiotic aspects
Type: aquarel ¹ (3) tekening ¹ (3)	Subject: Jogjakarta ¹ (4) Tweede politionele actie ¹ (7)
Technique: aquarelleren ¹ (3)	1948-12-19 ¹ (4) 1949-06-30 ¹ (1)
Material: hardboard ¹ (4)	militaire geschiedenis ¹ (12)

Associated Objects (25) [< prev](#) [1](#) [2](#) [3](#) [4](#) [5](#) [next >](#)

 President Soekarno g... Associated Press	 Sinkin panjang met s... Anonymous	 Indonesië vrij! Hatta, Mohammad	 Schild van een Ateher Anonymous	 Aankomst van Van Spi... Anonymous	 Het kasteel van Bata... Beeckman, Andries
--	---	---	---	--	---

Figure 1: Screenshot of object page in the Agora Event Browsing Demonstrator

4. THE AGORA DEMONSTRATOR

The automatically generated event thesaurus is applied in a new historical event browser called Agora² which provides an integrated access route to museum objects and audio-visual material from RMA and S&V respectively. It is a first step towards a platform to investigate the added value of historical events and narratives for the exploration of integrated collections. For each event and object there is an automatically generated page that shows (1) all associated objects, e.g., museum and audio-visual objects; (2) all associated events and the type of their relationship, e.g., previous-in-time event, sub-event; (3a) the event descriptive metadata, e.g., actors, place, period; or (3b) object descriptive metadata organized in three groups, e.g., biographical, material and semiotic dimensions – see figure 1 for a screenshot – and finally (4) the navigation path. The current version of the event thesaurus will be extended further to accommodate searching for relations between events such as temporal inclusion, causality and meronymy.

5. DISCUSSION

In this paper, we presented a modular pipeline for capturing knowledge about historical events from Dutch texts. Compared with previous approaches (i.e., [5]), it relies on a minimum of manual annotation and can be repurposed for other languages. To the best of our knowledge, this is the first work to extract events from unstructured Dutch text. Although our results are promising, more sophisticated techniques are necessary to obtain more fine-grained extractions and define measures for the historic relevance of the extracted events. Additionally, we also aim to find and represent relations between events such as causality, meronymy and correlation.

6. ACKNOWLEDGEMENTS

This research was funded by the CAMeRA Institute of the VU University Amsterdam and by the CATCH programme, NWO grant 640.004.801.

²<http://agora.cs.vu.nl/eventdemo>

7. ADDITIONAL AUTHORS

Lora Aroyo (VU University Amsterdam), Guus Schreiber (VU University Amsterdam) and Bob Wielinga (VU University Amsterdam), Jacco van Ossenbruggen (CWI and VU University Amsterdam), Johan Oomen (Netherlands Institute for Sound and Vision), Geertje Jacobs (Rijksmuseum Amsterdam).

8. REFERENCES

- [1] R. Cilibrasi and P. Vitanyi. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19(3):370–383, 2007.
- [2] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 2005.
- [3] G. Geleijnse, J. Korst, and V. de Boer. Instance classification using co-occurrences on the web. In *Proceedings of the ISWC 2006 workshop on Web Content Mining (WebConMine)*, Athens, GA, USA, November 2006.
- [4] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *ACM Events in MultiMedia Workshop (EiMM10)*, Oct 2010.
- [5] N. Ide and D. Woolner. Exploiting semantic web technologies for intelligent access to historical documents. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, pages 2177–2180, Lisbon, Portugal, 2004.
- [6] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI '99*, pages 474–479, 1999.
- [7] R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In *4th Annual Asian Semantic Web Conference (ASWC'09)*, 2009.
- [8] W. R. van Hage, V. Malaisé, G. de Vries, G. Schreiber, and M. van Someren. Abstracting and reasoning over ship trajectories and web data with the Simple Event Model (SEM). *Multimedia Tools and Applications*, 2011.