Dutch parliamentary debates on video

Marx, M.; Kooij, R.

Publication date
2010

Document Version
Final published version

Published in
Proceedings of the 10th Dutch-Belgian Information Retrieval Workshop (DIR 2010)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Dutch Parliamentary Debates on Video

Maarten Marx and Robert Kooij
ISLA, University of Amsterdam
Kruislaan 403 1098 SJ Amsterdam, The Netherlands
maartenmarx@uva.nl

ABSTRACT
We created an archive of video footage of the meetings of the Dutch parliament. All video was aligned with the official transcripts (a.k.a. Hansards or proceedings) of these meetings. A prototype search interface was built. The paper describes the data and the main technical aspects of the project.

Categories and Subject Descriptors
H.4.m [Information Systems]: Miscellaneous

Keywords
Dutch, Video, Politics

1. OBJECTIVE
Starting with the Scottish Parliament [5], several legislative bodies provide information systems giving access to video footage of their plenary meetings. The most advanced systems have segmented the video into natural units. Typically these are the topics discussed at the meeting, or, more fine-grained, the speeches made by members of the parliament or council. The same segmentation is made in the verbatim proceedings of the meeting and the two media are aligned.

This setup yields a richly annotated video data set for which entry point retrieval can be implemented using rather standard IR systems. Techniques from XML retrieval systems [3, 6] apply well for this use case. A well-designed example is the site http://theyworkforyou.com created by MySociety.

The objective of our work was to create a similar system for the Dutch Parliament. The availability of Dutch parliamentary proceedings is rather complicated:

- recent data (less than three weeks old) is available in HTML format from the website of the Dutch Parliament http://www.tweedekamer.nl
- video data is not archived and was at the time we started this research not systematically stored. Data is streamed live on the web at http://www.tweedekamer.nl

The PoliDocs parliamentary information system [1] provides a single access point to all parliamentary data in one uniform XML format. Proceedings are segmented into topics and blocks of speeches with interruptions. These blocks are divided into speeches by speakers, which are again divided into paragraphs. For every word being said in parliament, it is thus explicitly coded in the XML markup who said it, in what context, and when. An entry point retrieval system which returns speeches is available at http://polidocs.nl.

Our aim was to extend each search result in that system with a link to the exact point in the video of the meeting in which the found speech was made. A proof of concept was built and can be seen at http://openkamer.unwind.nl.

We briefly describe some of the technical aspects of this system. Figure 1 contains a screen shot.

Data format.
The video data is stored in two formats: Windows Media Video (source format) and Adobe Flash Video. The Flash Video format is obtained from the Windows Media Video using FFmpeg. The videos come with the following metadata: ID, date, time and duration. Through the ID each video is linked to the proceedings in XML format. Each speech in the XML file contains an attribute with a timecode.

Data collection.
We collected video data since 1 March 2009. At the time of writing (December 2009), the video corpus comprises 4190 hours, taken from 401 meetings (both plenary and committee) held at 117 days. The corpus occupies 838 GB of disc space.

Downloading video material from a webstream inevitably leads to some data loss. We use a script which quickly recovers after a break in the transmission. On average the script picks up the stream within less than a minute. We estimate that less than 4% of the transmitted video is lost. For a daily overview see http://www.openkamer.tv/status.php.
Aligning text and video.

The main problem in the project was the alignment of video and text. As we had segmented the text already at the level of speakers all we needed was to segment the video into speakers as well. We considered the following four techniques for this:

1. Obtain explicit timecodes from the stenographical section of the Dutch Parliament. This turned out to be impossible.

2. Every speaker change in the Dutch parliament involves a change in the used microphone (chairman, central lectern, and interruption microphone). Speaker segmentation would thus have been easy if we had access to the multi-track audio tape. This turned out to be impossible.

3. Speaker segmentation (also called diarization) using the audio and text files. We tried two systems but the results were not good enough for a live system.

4. Manual segmentation. We created a web based alignment tool with which the speaker segmentation and alignment of proceedings and video can be performed. The tool works smoothly. After a short training our coders could align 200 hours of video in 120 hours.

Acknowledgements

Maarten Marx acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement FOX, number FP7-ICT-233599.

De ontwikkeling van Openkamer.tv is mogelijk gemaakt door een bijdrage van Kennisland in de regeling Digitale Pioniers ronde eParticipatie (die door het ministerie van Binnenlandse Zaken en Koninkrijksrelaties genitieerd is). Thanks are due to Steven Grijzenhout and Ernst van Rheezen for providing scraper scripts.

2. REFERENCES