Semantic Dispatching of Multimedia News with MEWS

Julien Law-To
Exalead
10, place de la Madeleine
Paris, France
Julien.lawto@3ds.com

Gregory Grefenstette
Exalead
10, place de la Madeleine
Paris, France
gregory.grefenstette@3ds.com

Rémi Landais
Exalead
10, place de la Madeleine
Paris, France
Remi.landais@3ds.com

ABSTRACT
Online news comes in rich multimedia form: video, audio, photos, in addition to traditional text. Recent advances in semantically-rich text processing, in speech-to-text processing, and in image processing allows us to develop new ways of presenting and enriching news stories. Here we present MEWS, a Multimedia nEWS platform, which enriches news browsing according to media (text, images, and video) and to automatically detected type of news (music, general news, politics).

Categories and Subject Descriptors
H.5.1 Multimedia Information Systems; I.2.7 Natural Language Processing: *Speech recognition and synthesis*; I.4 Image Processing and Computer Vision

Keywords
Multimedia indexing, image processing, news, search engine, speech processing, speech-to-text, semantic annotation

1. INTRODUCTION
MEWS, a Multimedia nEWS search platform, applies latest advances in semantic text processing, in large-scale web indexing, in video, audio and image signal processing to provide an intelligent news browsing experience in which stories are enriched with domain specific metadata with technology developed by Exalead (crawling and indexing), Vocapia (speech processing), LTU (Image processing).

2. SOURCES TO ENRICH NEWS
Text-based new stories come from Exalead, the search engine division of Dassault Systèmes, which maintains a fresh index of 16 billion web pages, with 200 million reindexed per day, searchable at exalead.com/search. A separate index of web pages from news sources is also maintained, and a non-commercial demonstrator called ExaNews, provides access to 10 million of these news items at exanews.labs.exalead.com. The Exanews index feeds into MEWS, but MEWS also includes other text and multimedia news sources: podcasts, images, public debates, and less newsworthy items such as concert schedules of musicians.

Audio and video news sources come from the 300000 podcasts that have been indexed in the Exalead Voxalead [1] broadcast news search system, which has been live and online since 2009. Voxalead currently indexes daily broadcast news content from 60 sources in English, French, Chinese, Arabic, Spanish, Dutch, Italian, German and Russian and makes the audio and video streams searchable the day after their publication. The audio portions of news is converted from speech to time-stamped text using Vocapia technology [2]. This text is processed using named-entity recognition technology developed by Exalead, and entity and their types (events, people, organizations, places) are indexed.

Over 500,000 images are included in MEWS coming from the body of news gathered in the Exanews index, using Open Source solution Boilerpipe (https://code.google.com/p/boilerpipe) and HTML extraction to identify the largest image in each news story, ignoring small image, or image with aspect ratios resembling banners. These images are indexed by the text in any image tags and with text found around the image on the original web page.

In addition to news, MEWS also includes recent French National Assembly debate transcripts that have been provided as raw linked open data by data.gouv.fr. MEWS provides the first public index into this data, providing faceted, full text search, with further interface developments that exploit the structure of parliamentary debates. For some debates, we can also browse the video recording of these debates using an automatic alignment and synchronization performed by Vecsys between the manual transcript and the video [3].

To further enrich news results found via a user query, we also index all the pages of the English version of Wikipedia (about 10 million articles).

As many queries concern musical artists, we also incorporate in MEWS a music index concerning more than 600000 artists and their songs. Songs are indexed by textual metadata and lyrics but also by the musical content itself: chords, extraction of moods (happy, romantic etc.), of instrumentation (electric vs. acoustic). This index was produced using processing from ParisTelecom and the IRCAM, and is described in detail in [4].

3. SEARCHING, BROWSING IN MEWS
MEWS provides a unique search box provides access to all news. User queries can be automatically translated, using Systran technology, into another language (English, French, Spanish, Italian or German) to search. Results from all the different media sources (see section 2) are merged and displayed as shown below in Figure 1. Each result is illustrated with an image, subtitled with a search snippet and a distinguishing colored icon (showing whether the result is a video, an image, from the political domain, or from the music domain). If there is a match in the Wikipedia index, then a Wikipedia snippet appears below the search box, above the result images. The results scroll to the right (the fourth column in the image is the beginning of another three column result set). Clicking one of the search results leads to a detailed results page specific to this type of content.

MM'13, October 21-25, 2013, Barcelona, Spain.
ACM 978-1-4503-2404-5/13/10
http://dx.doi.org/10.1145/2502081.2502253

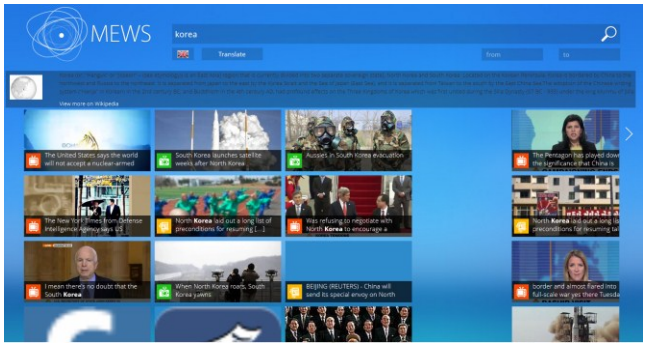

Figure 1. MEWS Search Page. Here the query 'Korea' produces results from Wikipedia, from text-based new stories, from automatic transcriptions of news broadcasts, and from tagged images.

3.1 Enriched display in detail page

An example of the detailed results page is shown in Figure 2. Terms from the original query are highlighted in bold in the results. The content and format of this news item depends of the type of document displaying either automatic transcription for videos (example shown here), manual transcription for politic content, tags for images or text for web pages. Tabs above the text allow the content viewed in English, French, Spanish, Italian and German, automatically translated using Systran technology. A tag cloud to the left of the content presents the main color-coded named entities detected in this content. Clicking on this tag cloud leads to Wikipedia articles for each entity.

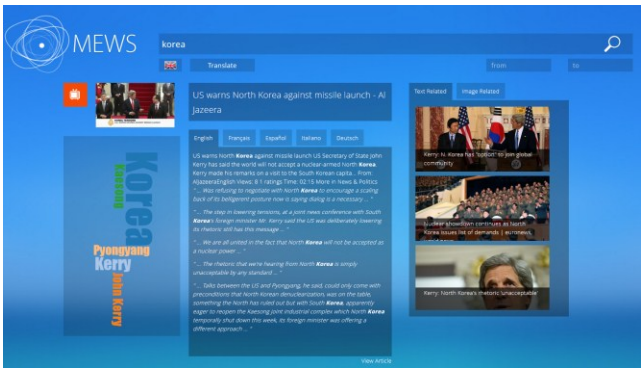

Figure 2. MEWS Detail Page, showing translated text tags, content with highlighted query terms, entity tag cloud and related content.

The content is also analyzed on the fly by an Exalead semantic processor that detects political people and artists. When an artist is detected, the named is highlighted in yellow and an automatic link leading to the artist page in our music index is created. In a similar way when a political person is detected a link to the political index demo is created. To the right of the central content, there is a column of similar content, found using keywords extracted from the news item. When the query result is an image, the similarity is based on the visual content using the image index and LTU's image comparison technology.

Figure 3. Voxalead Page accessible from MEWS. The video positions itself before the first mentions of the query terms. Small white triangles show the appearances of these terms under the video. The automatically produced transcript follows the video. Entities mentioned in the video appear under a search box on the right.

When the central content transcription comes from a video, clicking on the image icon will lead to the Voxalead page for this video, bringing the user directly to the moment in the video when the query terms are pronounced in the audio stream.

The live MEWS demonstrator, updated daily, is available at: http://mews.labs.exalead.com

4. ACKNOWLEDGMENT

This multimedia news demonstrator was partly realized as part of the Quaero Programme funded by OSEO, the French State Agency for Innovation.

5. REFERENCES

[1] Law-To, J., & Grefenstette, G. (2011, November). VOXALEAD: a scalable video search engine based on content. In Proceedings of the 19th ACM international conference on Multimedia (pp. 747-748). ACM.

[2] Despres, Julien, et al. "The Vocapia Research ASR Systems for Evalita 2011." Evaluation of Natural Language and Speech Tools for Italian. Springer Berlin Heidelberg, 2013. 286-294.

[3] Clavel, C., Adda, G., Cailliau, F., Garnier-Rizet, M., Cavet, A., Chapuis, G., ... & Suignard, P. (2013). Spontaneous speech and opinion detection: mining call-centre transcripts. Language Resources and Evaluation, 1-37

[4] Lenoir, A., Landais, R., & Law-To, J. (2012, June). MuMa: a scalable music search engine based on content analysis. In Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on (pp. 1-4). IEEE

[5] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.