Making Sense of the News

When is a news item news, and when does it change **my** world?
Primary contact: raphael.troncy@cwi.nl

1 Motivation

News web sites such as Le Monde[1], El Pais[2] or La Repubblica[3] generally classify news in categories such as: *World*, *National*, *Politics*, *Business*, *Science and Technology*, *Sport*, *Entertainment* and *Health*, while other services such as Google News[4] aggregate stories from multiple sources and offer personalised selections based on the user topics of interest. More advanced web sites such as SiloBreaker[5] or Newstin[6] provide more flexible access to news stories by *topic*, *person*, *organization* or *region*. These support the user's information need more closely, but just add more sources of individual news items, which leads to overly complex interfaces.

In existing news workflow processes, news items are accompanied by a fixed set of metadata descriptions that facilitate their storage and retrieval. These news items are typically *i)* produced by news agencies, independent journalists or citizen media, *ii)* consumed and enhanced by newspapers, magazines or broadcasters for finally *iii)* being delivered to end users. However, much of the metadata is lost because of interoperability problems occurring along the workflow. In addition at the end user interface, opportunities for making use of the metadata that is available are lost.

As a result, current systems have a number of limitations that leave the user to explore news information in an environment that contains **large amounts of irrelevant, unreliable and repeated information**. In particular, current systems:

- are unable to show how different news articles, images and videos are related to each other and are unable to provide explicit relationships between different sources on the same event to help the user form his/her own opinion on a particular topic;
- are unable to convey an overview of individual pieces of information and give insufficient support to make sense of individual news articles by providing links to background knowledge;
- are unable to provide a historic perspective of past events, e.g., highlighting "editorial" news items summarizing an event that took place some number of years ago.

[1] http://www.lemonde.fr/
[2] http://www.elpais.com/
[3] http://www.repubblica.it/
[4] http://news.google.com/
[5] http://www.silobreaker.com/
[6] http://www.newstin.com/

Consequently, users are overwhelmed by too many individual and disconnected pieces of information. Our hypothesis is that semantic processing of news information can improve the clustering and organization of individual news items – from heterogeneous sources, in multiple media types and in multiple languages – into meaningful events linked to appropriate background knowledge. In this project, we will build the technological infrastructure to allow the aggregation of multiple, distributed information sources. Based on this, we will provide user interfaces driven by semantic metadata for searching and browsing multimedia news articles, independently of whether the news is expressed in text or audio/visual media.

2 Project Goals

Our goal is to create an environment that facilitates end-users in seeing meaningful connections among individual news items (stories, photos, graphics, videos) through underlying knowledge of the descriptions of the items, their relationships and related background knowledge.

Enabling access to repositories containing any kind of media requires a system that can produce, collect, maintain and distribute media assets as well as aggregations of metadata associated with them. We will create metadata models to improve metadata interoperability along the entire news production chain. The underlying research challenges cover the two ends of the news workflow spectrum: how to model and represent semantic multimedia metadata along the news workflow and the consequences of this modeling at the user interface. At the same time, we will investigate the requirements the interface imposes on the modeling.

We will target two groups of users:

1. journalists: who require highly functional interfaces for searching for particular news items (from many sources, in several languages, on different media), and

2. lay users: who require simpler interfaces for being kept up-to-date on events of the day/week/month/year.

The high-level goal is to deploy semantic metadata throughout the workflow in specific tools, while at the same time using the requirements of the tools to optimise the metadata provided at several steps of the workflow. More specifically, we will deploy metadata to improve the detection of overlapping feeds of information to enable clustering into events the user is looking for. By coupling metadata with controlled vocabularies, these can be used for finding relations among items in different repositories. By linking the metadata with background knowledge available on the Web, such as Wikipedia[7] or Britannica[8] for information on people, events, countries and general topics, we will provide end users with access to contextual information to help with understanding individual news items. The controlled vocabularies and background knowledge

[7] http://www.wikipedia.org/

[8] http://www.britannica.com/nations

will in turn be used to “power” interfaces for end users to gain higher-level access to the repositories’ contents. We will create an environment where media assets can be enriched with information based on usage patterns. We will finally develop interface design guidelines, based on the metadata requirements, that are useful for different types of applications that manage multimedia assets.

While we need to automatically extract metadata from textual and visual resources, we do not claim to make any progress in textual or visual processing. Our hypothesis is that sufficient amounts of metadata are already available, or can be extracted with existing techniques. The problem is that metadata is lost along the workflow due to interoperability problems and/or that metadata is not used in the end-user application.

Using general data modeling and processing techniques we will:

- ensure interoperability along the news workflow by integrating existing knowledge models (e.g. the NAR ontology or COMM, the Core Ontology of Multimedia);
- find relations between news items coming from different sources, in different languages and on different media by using these models;
- cluster news items by using statistical analyses;
- rank the items of a cluster and find the representative of this cluster by analyzing the structure of the knowledge base.

In addition, we will use standard HCI methods (e.g. task analysis, user centred design) to ensure that the interface follows the maxims of conversation of Paul Grice[9] by:

- rendering an event and its related news items;
- providing the context needed to interpret a news item by providing links to related background knowledge at an appropriate level of detail;
- providing appropriate levels of information granularity, for example, a global view linking to more detailed information.

3 Exploitation

AFP and VRT will exploit the results of the project in their daily business activity. The metadata integration will improve the overall production chain: broadcasters will produce TV news more easily based on the intelligent processing of news stories from news agencies, while archivists will have, for free, rich descriptions of media assets and thus be able to search and reuse easily old news content.

News content production is characterized by a dynamic and flexible process where actors such as AFP and VRT play a prominent role. As opposed to traditional workflows, where a news agency such as AFP would provide raw material and a news producer such as VRT would take care of the packaging and distribution to the consumer, we are now in a situation where these roles

[9] `http://en.wikipedia.org/wiki/Paul_Grice#Conversational_Maxims`

have become interchangeable. Using internet technology, news agencies have direct access to the end user and through the European Broadcasting Union (EBU), traditional broadcasters have in fact become part of a news agency network. The net result of this will be that news agencies, broadcasters or news publishers in general, as well as the end user are confronted with a massive increase of available information.

The designed infrastructure will provide journalists with extraction of concepts and metadata, suggestion of related or similar news, selection of multimedia objects (photos, videos, sounds) related to the news they are currently working on, in order to produce richer and more relevant content. It will also provide tools to link different media objects between them in a semi or automatic way, based on the metadata. Journalists will be thus guided for producing faster and better multimedia news content by optimizing their tasks such as searching for particular facts, gathering information on a given topic, finding quotes in a video, etc.

Beyond these two particular companies, the whole sector (news agencies, independent journalists and broadcasters) will benefit from these technologies because we will work in close collaboration with the International Press Telecommunication Council[10] (IPTC) gathering all news agencies in the world through AFP, the leader of the metadata working group in this standardization body.

We will also collaborate actively with the European Broadcaster Union[11] (EBU), which has set up a particular programme in the context of news production automation, through VRT, that is committed to deliver the reference software for this programme, and through CWI, that acts as consultant for the future Electronic Program Guides development.

We will disseminate our work by continuing our activities within the W3C[12] consortium, and particularly in the forthcoming working group that will produce the standards to make video a first class citizen in the future web.

4 Relationship to Existing Projects

NEWS[13] developed an automated multi-lingual textual news classification and annotation engine that is able to solve ambiguities and find matching events. We will be able to build on the results of this work to facilitate clustering news items into closely related groups. We will also tackle multiple types of media, in addition to text, as it is our belief that video will gradually become the dominant medium for news content.

Citizen Media[14] enables lay users to consume, author and publish their content as part of a networked audiovisual system. The project focuses on automatic analysis of visual information and on scalability issues since the system has to be able to handle a massive amount of user-generated content in different formats in real time, and annotate and store this content in huge databases in order to better reuse all these pieces of user-generated content. We will build on the experiences with the interfaces developed for lay users but in addition

[10] http://www.iptc.org/

[11] http://www.ebu.ch/

[12] http://www.w3.org/

[13] http://www.news-project.com/

[14] http://www.ist-citizenmedia.org/

will target professional users from news agencies and broadcasters as well as independent journalists. We will also emphasize the role of semantic metadata for solving interoperability problems and empowering end-user interfaces.

MESH[15] aims to extract, compare and combine content from multiple multimedia news sources, automatically create advanced personalized multimedia summaries, syndicate summaries and content based on the extracted semantics, and provide end users with a *multimedia mesh* news navigation system. While the project has made progress across this broad set of goals, it focuses mainly on news distribution on mobiles. We concentrate, however, on the use of the underlying technological semantic infrastructure to reduce the amount of information exposed to the user in a simplified interface.

5 Why us?

The consortium is an interdisciplinary team of engineers and researchers in Computing Science and Web Science, journalists and professionals in the media industry. Realizing this ambitious vision requires the participation of a major news agency provider (AFP) and a news agency consumer that provides broadcast news (VRT) to cover the entire news production workflow. Automatic processing of the multimedia news content requires an expert in image and video analysis (IBBT) and an expert in textual news content analysis (Temis, the European leader in language processing that is able to deal with 20 languages). Finally, the consortium needs an integrator that understand user needs and is expert in using semantic web technologies to describe multimedia news content (CWI).

[15] `http://www.mesh-ip.eu/?Page=Project`