TRENTINOMEDIA: Exploiting NLP and Background Knowledge to Browse a Large Multimedia News Store*

Roldano Cattoni[1], Francesco Corcoglioniti[1,2], Christian Girardi[1], Bernardo Magnini[1], Luciano Serafini[1], and Roberto Zanoli[1]

[1] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[2] DISI, University of Trento, Via Sommarive 14, 38123 Trento, Italy

{cattoni,corcoglio,cgirardi,magnini,serafini,zanoli}@fbk.eu

Abstract. TRENTINOMEDIA provides access to a large and daily updated repository of multimedia news in the Trentino Province. Natural Language Processing (NLP) techniques are used to automatically extract knowledge from news, which is then integrated with background knowledge from (Semantic) Web resources and exploited to enable two interaction mechanisms that ease information access: entity-based search and contextualized semantic enrichment. TRENTINOMEDIA is a real multimodal archive of public knowledge in Trentino and shows the potential of linking knowledge and multimedia and applying NLP on a large scale.

1 Introduction

Finding information about an entity in a large news collection using standard keyword-based search may be time-consuming. Searching for a specific person, for instance, may return a large list of news about homonymous persons, that need to be checked and filtered manually. Also, understanding the contents of a news can be expensive, if the user is not familiar with the entities mentioned and needs information about them.

The presented TRENTINOMEDIA system shows how the use of NLP and Semantic Web techniques may help in addressing these problems. TRENTINOMEDIA supports the "smart" access to a large and dynamic (daily updated) repository of multimedia news in the Italian Trentino Province. "Smart" means that NLP techniques are used to automatically extract knowledge about the entities mentioned in the news. Extracted knowledge is then integrated with *background knowledge* about the same entities gathered from (Semantic) Web resources, so to build a comprehensive knowledge base of entity descriptions linked to the news of the collection. Exploiting the interlinking of knowledge and multimedia, two interaction mechanisms are provided to ease information access:

- *entity-based search*, enabling a user to find exactly the news about a specific entity;
- *contextualized semantic enrichment*, consisting in the visualization of additional knowledge about a mentioned entity that may ease a user's understanding of a news.

Two main usages are foreseen for TRENTINOMEDIA: (i) a professional usage, restricted to the news providers and aimed at addressing internal needs, including automatic news documentation, support tools for journalists and integration of advanced

* This work was supported by the LiveMemories project (http://www.livememories.org) funded by the Autonomous Province of Trento (Italy) under the call "Major Project".

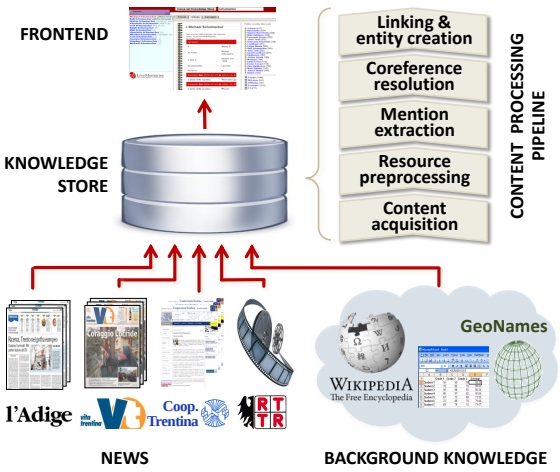

Fig. 1: System architecture

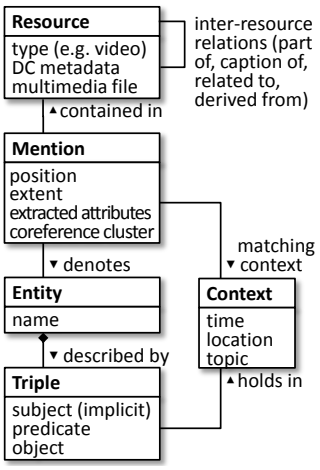

Fig. 2: Data model

functionalities in existing editorial platforms; and (ii) an open use by citizens through on-line subscriptions to news services, possibly delivered to mobile devices.

The presented work has been carried out within the LiveMemories project, aimed at automatically interpreting heterogeneous multimedia resources, transforming them into "active memories". The remainder of the paper presents the system architecture in section 2 and the demonstrated user interface in section 3, while section 4 concludes.

2 System architecture

The architecture of TRENTINOMEDIA is shown in figure 1 and includes three components: the KNOWLEDGESTORE, a *content processing pipeline* and a *Web frontend*.

The KNOWLEDGESTORE [3] builds on Hadoop[3] and Hbase[4] to provide a scalable storage for multimedia news and background knowledge, which are represented according to the (simplified) schema of figure 2. News are stored as multimedia *resources*, which include texts, images and videos. Knowledge is stored as a *contextualized ontology*. It consists of a set of *entities* (e.g., "Michael Schumacher") which are described by ⟨subject, predicate, object⟩ RDF [2] *triples* (e.g., ⟨Michael Schumacher, pilot of, Mercedes GP⟩). In turn, each triple is associated to the ⟨time, space, topic⟩ *context* the represented fact holds in (e.g., ⟨2012, World, Formula 1⟩). The representation of contexts follows the Contextualized Knowledge Repository approach [5] and permits to accommodate "conflicting" knowledge holding under different circumstances (e.g. the fact that Schumacher raced for different teams). Resources and entities are linked by *mentions*, i.e. proper names in a news that refer to an entity. They permit to navigate from a news to its mentioned entities and back, realizing the tight interlinking of knowledge and multimedia at the basis of the interaction mechanisms of TRENTINOMEDIA.

Concerning the content processing pipeline, it integrates a number of NLP tools to load, process and interlink news and background knowledge, resulting in the full

[3] http://hadoop.apache.org
[4] http://hbase.apache.org

Table 1: Resource statistics

Provider	News	Images	Videos
l'Adige	733,738	21,525	-
VitaTrentina	33,403	14,198	-
RTTR	2,455	-	120 h
Fed. Coop.	1,402	-	-
Total	770,998	35,723	120 h

Table 2: Extraction, coref. and linking stats.

Entity type	Recognized mentions	Mention clusters	Linked clusters	Total entities
persons	5,566,174	340,147	5.03%	351,713
organiz.	3,230,007	16,649	7.96%	17,129
locations	3,224,539	52,478	48.64%	52,478
Total	12,020,720	409,274	10.74%	421,320

population of the schema in figure 2. Apart from the loading of background knowledge, which is bootstrapped by manually selecting and wrapping the relevant knowledge sources, the pipeline works automatically and incrementally, processing news as they are collected daily. The rest of this section describes the processing steps of the pipeline, while the user interface of the frontend is described in the next section.

Content acquisition. News are supplied daily by a number of news providers local to the Trentino region. They are in Italian, cover a time period from 1999 to 2011 and consist of text articles, images and videos. Loading of news is performed automatically and table 1 shows some statistics about the news collected so far. Background knowledge is collected manually through a set of ad-hoc wrappers from selected (Semantic) Web sources, including selected pages of the Italian Wikipedia, sport-related community sites and the sites of local and national public administrations and government bodies. Overall, it consists of 352,244 facts about 28,687 persons and 1,806 organizations.

Resource preprocessing. Several operations are performed on stored news with the goal of easing their further processing in the pipeline. Preprocessing includes the extraction of speech transcription from audio and video news, the annotation of news with a number of linguistic taggers (e.g., part of speech tagging and temporal expression recognition, performed using the TextPro tool suite[5] [6]) and the segmentation of complex news in their components (e.g., the separation of individual stories in a news broadcast or the extraction of texts, figures and captions from a complex XML article).

Mention extraction. Textual news are processed with TextPro to recognize mentions of three types of named entities: persons, organizations and geo-political / location entities. For each mention, a number of attributes is extracted from the mention and its surrounding text. Given the text "the German pilot Michael Schumacher", for instance, the system recognizes "Michael Schumacher" as a person mention and annotates it with FIRSTNAME "Michael", LASTNAME "Schumacher", ROLE "pilot" and NATIONALITY "German". Attributes are extracted based on a set of rules (e.g., to split first and last names) and language-specific lists of words (e.g., for nationalities, roles, ...). Statistics about the mentions recognized so far are reported in the second column of table 2.

Coreference resolution. This step consists in grouping together in a *mention cluster* all the mentions that (are assumed to) refer to the same entity, e.g., to decide that mentions "Michael Schumacher" and "Schumi" in different news denote the same person. Two coreference resolution systems are used. Person and organization mentions are processed with JQT2 [8], a system based on the Quality Threshold algorithm [4] that

[5] http://textpro.fbk.eu

compares every pair of mentions and decides for coreference if their similarity score is above a certain *dynamic threshold*; similarity is computed based on a rich set of features (e.g., mention attributes and nearby words), while the threshold is higher for ambiguous names, requiring more "evidence" to assume coreference. Location mentions are processed with GeoCoder [1], a system based on geometric methods and on the idea that locations in the same news are likely to be close one to another; it exploits the Google Maps geo-referencing service[6] and the GeoNames geographical database[7]. Statistics about the mention clusters identified so far are reported in the third column of table 2.

Linking and entity creation. The last step consists in linking mention clusters to entities in the background knowledge and to external knowledge sources. Clusters of location mentions are already linked to well-known GeoNames toponyms by GeoCoder. Clusters of person and organization mentions are linked to entities in the background knowledge by exploiting the representation of contexts. The algorithm [7] firstly identifies the ⟨time, space, topic⟩ contexts most appropriate for a mention cluster among the ones in the KNOWLEDGESTORE, based on the mentions attributes and the metadata of the containing news (e.g., the publication date). Then, it searches for a matching entity only in those contexts, improving disambiguation. The fourth column of table 2 reports the fraction of mention clusters linked by the systems, i.e. the *linking coverage*: coverage is low for clusters (10.74%), but increases in terms of mentions (31.03%), meaning that the background knowledge mainly consists of popular (and thus frequently mentioned) entities. New entities are then created and stored for unlinked mention clusters, as they denote real-world entities unknown in the background knowledge; the last column of table 2 reports the total number of entities obtained so far. All the entities are finally associated to the corresponding Wikipedia pages using the WikiMachine tool[8].

3 User Interface

The entry point of the TRENTINOMEDIA Web interface is a search page supporting *entity-based search*. The user supplies a proper name which is looked up among the entities in the KNOWLEDGESTORE and a list of matching entities is returned for disambiguation; entities are listed by type and distinguished with short labels generated from stored information, as in figure 3, left side. By selecting an entity, the user is presented with the list of news mentioning that entity, retrieved based on the associations between entities, mentions and resources stored in the KNOWLEDGESTORE. A descriptive card is also displayed, as shown in the right side of figure 3. It contains all the information known about the entity, including: (i) background knowledge, (ii) information carried by the attributes of the entity mentions and (iii) frequently co-occurring and likely related entities. The example in figure 3 shows the potential but also the weaknesses of processing noisy, real world data with automatic NLP tools. In particular, typos and the use of different names for the same entity (e.g., acronyms, abbreviations) may cause coreference resolution to fail and identify multiple entities in place of one, as happens with "F1" and "Formula 1", "Raikkonen" and "Kimi Raikkonen", "Mich*ea*l Schumacher" and "Mich*ae*l Schumacher". Still, the use of additional information ex-

[6] http://code.google.com/apis/maps/
[7] http://www.geonames.org/
[8] http://thewikimachine.fbk.eu/

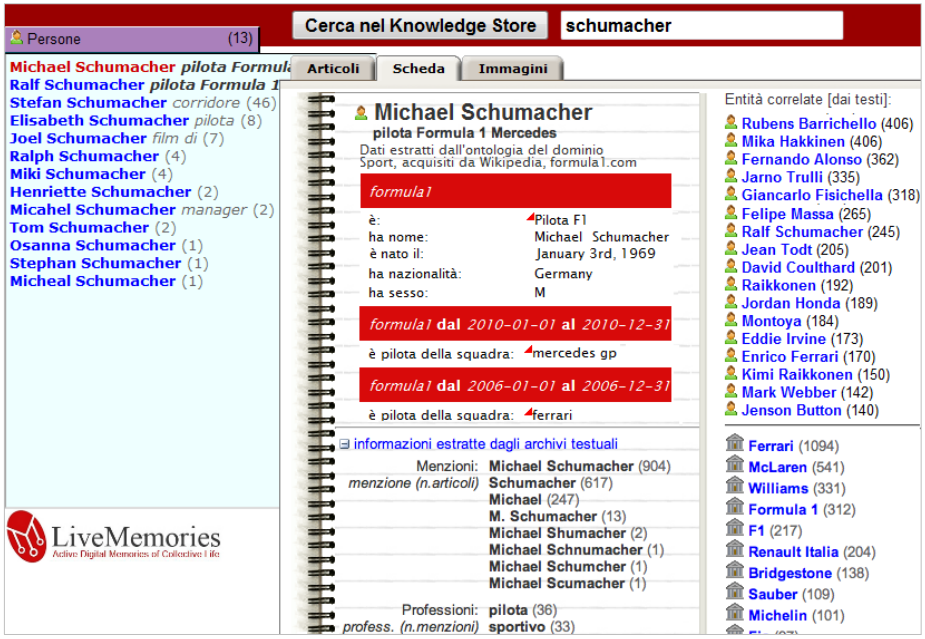

Fig. 3: Example of entity-based search for query "Schumacher" in TRENTINOMEDIA.

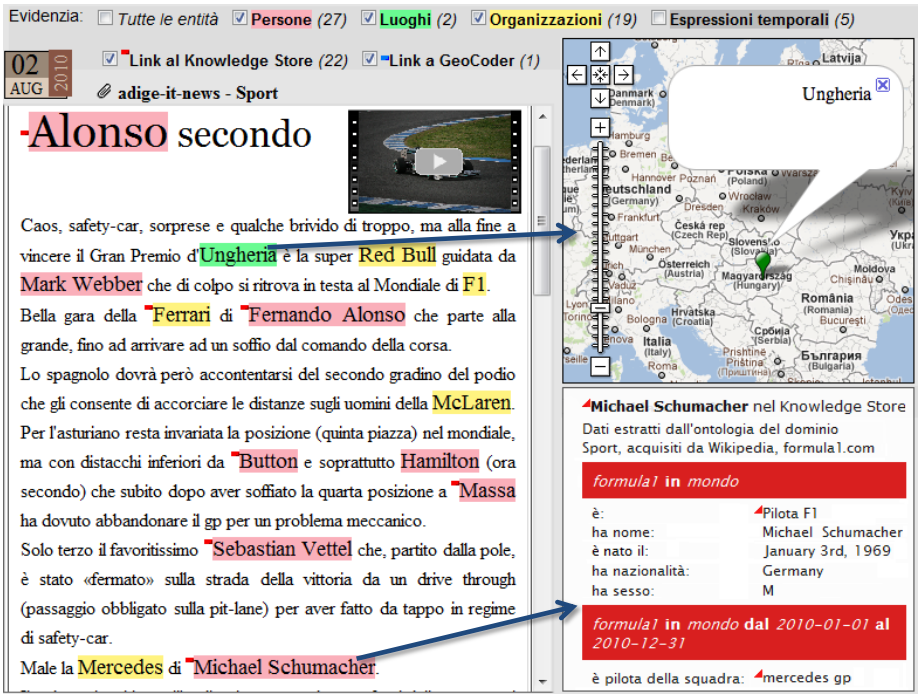

Fig. 4: SmartReader showing a sport news.

tracted from texts (e.g., keywords) can often overcome the problem, as happens with the correct coreference of "Schumacher", "Michael" and "Michael S*hu*macher").

The other interaction mechanism supported by TRENTINOMEDIA—*contextual semantic enrichment*—is accessed through the *SmartReader* interface shown in figure 4, which is displayed by selecting a news. The SmartReader allows a user to read news or watch videos while gaining access to related information linked in the KNOWLEDGESTORE to the news and its mentioned entities. The interface is organized in two panels. The left panel displays the text of the news or the video with its speech transcription and permits to selectively highlight the recognized mentions of named entities. The right panel provides *contextual information* that enriches the news or a selected mention. It can display a cloud of automatically extracted keywords, each providing access to related news. It can also show additional information about the selected mention, by presenting: (i) the Wikipedia page associated to the mentioned entity, (ii) a map displaying a location entity and (iii) a descriptive card with information about the entity in the background knowledge. In the latter case, only facts which are valid in the ⟨time, space, topic⟩ context of the news are shown, e.g., that "Schumacher is a pilot of Mercedes GP in 2010", so to avoid to overload and confound the user with irrelevant information.

4 Conclusions

TRENTINOMEDIA shows how the application of NLP techniques and the interlinking of knowledge and multimedia resources can be beneficial to users accessing information contents. In particular, two mechanisms to exploit this interlinking are demonstrated: entity-based search exploits links from knowledge (entities) to resources, while semantic enrichment exploits links in the opposite direction. TRENTINOMEDIA also shows that NLP and Semantic Web technologies are mature enough to support the large scale extraction, storage and processing of knowledge from multimedia resources.

References

1. Buscaldi, D., Magnini, B.: Grounding toponyms in an Italian local news corpus. In: Proc. of 6th Workshop on Geographic Information Retrieval. pp. 15:1–15:5. GIR '10 (2010)
2. Carroll, J.J., Klyne, G.: Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation (2004), http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
3. Cattoni, R., Corcoglioniti, F., Girardi, C., Magnini, B., Serafini, L., Zanoli, R.: The KNOWLEDGESTORE: an entity-based storage system. In: Proc. of 8th Int. Conf. on Language Resources and Evaluation. LREC '12 (2012)
4. Heyer, L.J., Kruglyak, S., Yooseph, S.: Exploring expression data: Identification and analysis of coexpressed genes. Genome Research 9(11), 1106–1115 (1999)
5. Homola, M., Tamilin, A., Serafini, L.: Modeling contextualized knowledge. In: Proc. of 2nd Int. Workshop on Context, Information And Ontologies. CIAO '10, vol. 626 (2010)
6. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proc. of 6th Int. Conf. on Language Resources and Evaluation. LREC '08 (2008)
7. Tamilin, A., Magnini, B., Serafini, L.: Leveraging entity linking by contextualized background knowledge: A case study for news domain in Italian. In: Proc. of 6th Workshop on Semantic Web Applications and Perspectives. SWAP '10 (2010)
8. Zanoli, R., Corcoglioniti, F., Girardi, C.: Exploiting background knowledge for clustering person names. In: Proc. of Evalita 2011 – Evaluation of NLP and Speech Tools for Italian (2012), to appear