RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News

Conference Paper *in* Lecture Notes in Computer Science · November 2005

DOI: 10.1007/11574620_10 · Source: DBLP

CITATIONS
19

READS
50

3 authors:

Roberto Basili
University of Rome Tor Vergata
252 PUBLICATIONS **2,386** CITATIONS
SEE PROFILE

Marco Cammisa
University of Rome Tor Vergata
20 PUBLICATIONS **171** CITATIONS
SEE PROFILE

Emanuele Donati
Abbott Diabetes Care
2 PUBLICATIONS **19** CITATIONS
SEE PROFILE

RitroveRAI: A Web Application for Semantic Indexing and Hyperlinking of Multimedia News*

Roberto Basili, Marco Cammisa, and Emanuale Donati

University of Roma, Tor Vergata, Department of Computer Science,
Via del Politecnico snc, 00133, Roma
{basili, cammisa, donati}@info.uniroma2.it

Abstract. In this paper, a system, RitroveRAI, addressing the general problem of enriching a multimedia news stream with semantic metadata is presented. News metadata here are explicitly derived from transcribed sentences or implicitly expressed into a topical category automatically detected. The enrichment process is accomplished by searching the same news expressed by different agencies reachable over the Web. Metadata extraction from the alternative sources (i.e. Web pages) is similarly applied and finally integration of the sources (according to some heuristic of pertinence) is carried out. Performance evaluation of the current system prototype has been carried out on a large scale. It confirms the viability of the RitroveRAI approach for realistic (i.e. 24 hours) applications and continuous monitoring and metadata extraction from multimedia news data.

1 Introduction

Web services actually tend to offer functional and non-functional requirements and capabilities in an agreed, machine-readable format. The target is the support to automated service discovery, selection and binding as a native capability of middleware and applications. However, major limitations are due to the lack of clear and processable semantics. Multimedia data are even more critical as semantics often depends on multiple and independent aspects: *functional information*, e.g. data format and processing constraints, *application criteria*, e.g. the different commercial constraints that may be applied, as well as *content information*, e.g. the *topics* to which a TV program refer or the *genre* of a song or video clip. In particular audio-visual data suffer from the fact that they are particularly rich in content and the level of semantic description is not easily detected from the different co-operating information (the video content vs. the environment sound as well as speaker's comments) that give rise to a variety of abstraction levels.

Methods of Information Extraction from multimedia data have thus to face specific problems in order to support realistic Semantic Web scenarios:

* The research work presented in this paper has been partially funded by the PrestoSpace IST Integrated Project, n. IST-FP6-507336.

Y. Gil et al. (Eds.): ISWC 2005, LNCS 3729, pp. 97 – 111, 2005.

- They must capture *levels of abstraction* able to express content at the visual level as well as at the sound (or speech) level
- Given the richness of the audio-visual information and the usually large size of target archives they must be *efficient* and *scalable*
- They should be as much *adaptable* as possible even in the early development phases in order to afford problems of realistic size. In particular, methods of machine learning for the construction of the required large knowledge bases and rule sets are needed.
- They must be robust with respect to noise and complexity (often incompleteness) of the source data

For the above reasons, extraction of audio-visual semantic metadata is thus a critical problem for a large class of Semantic Web applications. In the PrestoSpace[1] project (IST Integrated Project, n. IST-FP6-507336) the problem of preservation of the huge archives of European audio-visual providers (i.e. BBC, RAI and INA) through systematic digitalisation and restoration techniques has been pursued. In this scenario, the need for making digitised data accessible through intelligent information retrieval interfaces has been approached by the automation of semantic metadata extraction from raw material. Ontological resources are thus also used as a reference model for extraction and ontology-based and multilingual retrieval.

In this paper a system, RitroveRAI, developed for the semantic metadata extraction from TV and radio broadcasted news, is presented. It realizes the semantic extraction component of the overall PrestoSpace solution to preservation and indexing of audio-visual material: it is actually implemented for Italian over the data of the RAI TV channel. The RitroveRAI system makes use of human language technologies for IE over multimedia data (i.e. speech recognition and grammatical analysis of incoming news). News are topically categorised by means of a statistical categorizer. IE results are then exploited to find on the Web texts/pages equivalent (or weakly equivalent) to source news: this aims to extend the metadata derived from news with systematic material available on the aligned Web texts. Finally integration of internal (i.e. expressed by the source news) as well as external (as found on the Web) information produces the final set of metadata published with the digitised news.

The next section will introduce the overall approach by discussion the architecture of the current RitroveRAI prototype. Then advanced aspects like automatic (machine learning driven) categorisation of news, enrichment of broadcasted news via Web alignment and mining, will be discussed. Finally, performance evaluation results over large data sets will be reported to drive the final discussion.

2 An Approach to Metadata Indexing Based on HLT and the Web

The source information in RitroveRAI system have a well identifiable topic related to the content, i.e. the events and participants to which the reported news refer. Methods of extraction can use some visual information (in TV news) but are mainly tight to the

[1] URL: `http://www.prestospace.org/index.en.html`

speaker output. This is captured by a speech recognition system that initiates the processing chain. Some preprocessing focuses on the segmentation of source programs into individual news relying on special content features like the time span of silence intervals, detected changes in the speaker's voice or program schemes (e.g. alternating videos and studio contributions). The approach discussed in this paper refers to all the remaining Information Extraction and Web mining steps up to the final metadata publishing and browsing facilities. Figure 1 reports the overall client-server architecture.

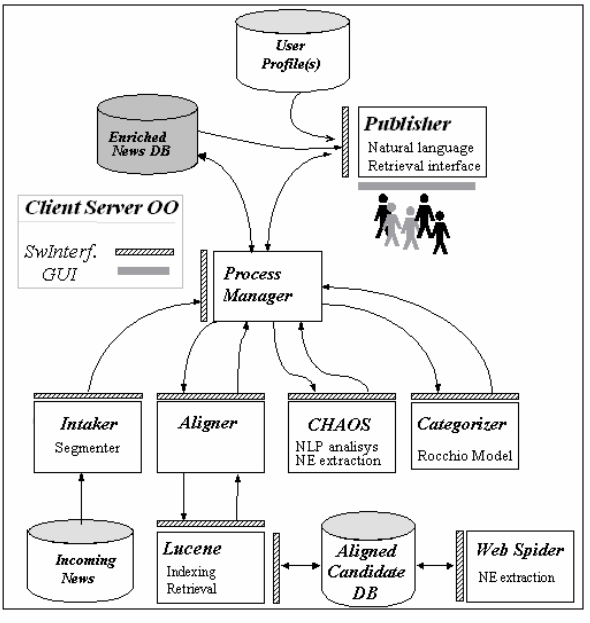

Fig. 1. The RitroveRAI Client-Server Architecture

The workflow processing is organized around a modular client-server architecture coordinated by a glueware module, called the "*Process Manager*". A second independent server is the *Web Spider*. It is based on the Google's API and retrieves all the documents published by a reference content provider (i.e. a journal) within a time window centered on the day of the news transmission. Finally a linguistic processing server, called *CHAOS* [2,4], makes available grammatical recognition capabilities over both transcriptions and Web pages.

The Information Extraction chain applies first the *Intaker* module. It collects and normalizes the incoming broadcasted news as they are transcribed and segmented by the speech recognition tool. The result of the intaking process is the update of news and segments into an internal DB structure responsible of supporting all the later processing stages. Then, the *Categorization* module is invoked by the Process Manager over the intaken news: it returns the pertinent topical categories (with their associated confidences) according to the RAI internal classification scheme. Concurrently, the *Aligner* module selects the candidates equivalent news from those extracted by the *Web Spider* process. This starts from the transcriptions parsed by the

CHAOS server. Web pages are also parsed[2] and indexed according to traditional IR models. For each news item, the *Alignment* process selects the Web pages from a set of candidate ones, i.e. those made available by the Web Spider, and create direct hyperlinks to them. This also allows to include auxiliary more precise metadata as those associated to the aligned news in order to prune the possibly irrelevant information[3]. Whenever internal and external metadata are made available, customized browsing (i.e. navigation through a user-specific hypertext built over the processed news) is allowed by a specialized *Web browsing interface*. Queries in Natural Language are also supported.

3 Natural Language Processing of Broadcasted and Web News

Natural Language processing is required in RitroveRAI for two purposes. First it enables the extraction from the source speech transcription of a number of phenomena: common nouns, verbs and Named Entities (e.g. person, location and organization names). Second, it also derives semantic information from the Web aligned news. However, these latter, being them written in plain natural language (i.e. not automatically and noisy transcribed) allow the extraction of grammatical relations: for example subject or object relation between named entities and verb. Notice that this has an impact on the system knowledge about the role played by individuals as participants to the target facts of the news.

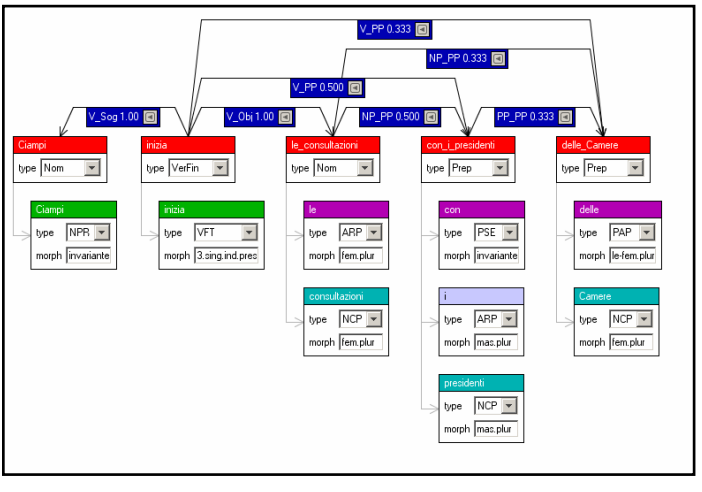

Fig 2. The dependency graph of an Italian sentence

[2] The parsing process is different in the two cases as automatic transcriptions follow less rigidly linguistic well-formedness criteria so that specific grammatical and lexical rules are required.

[3] When mistakes made by the speech recognizer over incoming transcriptions affect the quality of the source metadata, external, i.e. Web originated, metadata can be used to validate the former and compensate such errors.

Linguistic extraction is carried out by a Natural language parser called CHAOS [2,4]. Chaos is a server for modular and lexicalized parsing based on a cascade of morphosyntactic recognisers. The main CHAOS modules are: the *Tokenizer*, the *morphological analyzer* (who identifies the possible morphological interpretation of every token), a *part of speech tagger*, a *named entity recognizer*, a *chunker* (who collects possibly multiple tokens to form bigger grammatical and unambiguous units called *chunks*), the *temporal expressions recognizer*, a *verb subcategorization analyzer* (for the recognition of the main verbal dependencies) and a *shallow syntactic analyzer* (for the recognition of remaining and possibly ambiguous dependency, e.g. prepositional modifiers of verbs and nouns).

In Fig. 2, as an example, the dependency graph (called XDG) for the Italian sentence

"*Ciampi inizia le consultazioni con i presidenti delle camere*"[4]

is reported.

The *eXtended Dependency Graph* (*XDG*) formalism, introduced in [4], represents, the recognized sentence as a planar graph, whose nodes are constituents and arcs the grammatical relationships between them. The constituents are chunks, i.e. kernels of verb phrases (VPK), noun phrases (NPK), prepositional phrases (PPK) and adjectival phrases (ADJK). Examples of PPK and NPK chunks in Fig. 2 are respectively "*con i presidenti*"[5] and "*le consultazioni*"[6]. Relations among the constituents represent grammatical functions among their syntactic heads: logical subject (*lsubj*), logical objects (*lobj*), and prepositional modifiers. More technical details on the CHAOS parser can be found in [4].

3.1 Two Parsing Models for RitroveRAI

The incoming broadcast news set, hereafter called N_I, and the published Web news set, hereafter N_E, must be distinguished, as introduced above, to apply different parsing strategies.

Several difficulties arise when Chaos is applied to the N_I set, due to its noisy nature. First, frequent misspellings characterize elements in N_I. This is very variable depending on the speed, clarity, pronounce of the speaker and by the quality of the signal.

Another problem is that all the misspellings are also correct words of the Italian language. This can lead to errors in the named entities extraction, because semantically odd entities are also introduced in the text.

A third problem is the absence of case information in the transcribed text. Most technologies for NE extraction are actually based on capitalized words.

Finally, text segments are generated automatically during the transcription process. Speeches are translated into "news units" by using time and intensity rules. When the audio signal goes under a specified lower intensity for a sufficient time interval a new

[4] "*Ciampi starts the counsels with the presidents of the Chambers*".
[5] "*with the presidents*".
[6] "*the counsels*".

segment is initiated. These heuristics are not error free so that multiple news items may appear in a segment or, dually, a single news can be split into multiple (but less complete) segments. The segmentation process is a special case of the recognition of boundaries between distinct textual units in documents. Errors in this phase impact on the accuracy of all later processing steps in RitroveRAI.

CHAOS has been applied over the broadcasted news only to recognize a subset of grammatical data as wrong POS tagging is the general case: contexts are not reliable enough to trigger POS tagging rules. Here the recognition of basic distinctive information to support categorization and Web alignment is carried out. The transcription's parsing model supports shallow parsing including only the tokenizer and the morphological analyzer based on gazetteer lookup. Evaluation of the adopted design choice (Section "*Performance Evaluation*") confirms that such limited (but reliable) information is sufficient most of the times.

On the other hand, the Chaos full parsing cascade was applied with its full functionalities to the aligned Web news, i.e. the set N_E. This allowed to extract named entities as well as all their verbal relations from the N_E set. In this way Named Entities of the source news are confirmed (as they also appear in the aligned news found in the Web) but their role in the described fact is also captured most of the times. In the previous example we would know that (*Carlo Azeglio*) "*Ciampi*" (current President of the Italian Republic) is the `agent` initiating the counsels. This results in an higher abstraction level in the derivation of content metadata able to match more specific queries in future retrieval scenarios.

3.2 Recognizing Named Entities from Broadcasted News (NE_I set)

Named Entities in the incoming transcribed segments are herefter called internal named entities , i.e. NE_I. As an example, in the following segment "*calling for democracy and freedom, the leaders of iraq's interim government today challenged the country's new national assembly to strive for unity president bush congratulated the people of iraq. "it was a hopeful moment", bush told reporters at the white house*", the following NE_I list is derived: "*Iraq*", "*National Assembly*", "*Bush*", "*White House*". The gazetteers used by CHAOS are here used as a major source of information.

As a segment is to be categorized and then also aligned with other Web pages, it is also useful to recognize common nouns in the text. In the above example "*democracy*", "*leaders*", "*freedom*", "*government*", "*country*", "*people*" and "*reporters*" would be extracted. NE_I and other nouns are a surrogate of the segment transcription useful for categorization and Web mining. We make use of this information to build an efficient search vector for Web retrieval. In order to distinguish the different importance of Named Entities and common nouns, we modified slightly the usual weighting scheme of the IR platform adopted (i.e. Lucene [5]). Common nouns n are given a weight equal to their document frequency (occ_n) (i.e. the default weight in Lucene syntax). Named Entities are instead amplified by a factor w, with a resulting weight of $w \cdot occ_n$. We found that different domains require different ratios w. In all our settings $w=4$ is used.

4 Machine Learning for Broadcasted News Categorization

Text categorization is a traditional supervised machine learning task. In RitroveRAI a the Rocchio model, as a profile based classifier, presented in [3], has been used. Given the set of training document R_i, classified under the topics C_i (positive examples), the set $\overline{R}_i$ of the documents not belonging to C_i (negative examples) and given a document d_h and a feature f, the Rocchio model [8,3] defines the weight Ω_f of f in the profile of C_i as:

$$\Omega_f^i = \max\left\{0, \frac{\beta}{|R_i|}\sum_{d_h \in R_i} \omega_f^h - \frac{\gamma}{|\overline{R}_i|}\sum_{d_h \in \overline{R}_i} \omega_f^h\right\} \quad (1)$$

where ω_f^h is the weight of the feature f in the document d_h. In formula (1), the parameters β and γ control the relative impact of positive and negative examples and determine the weight of f in the *i-th* profile. In [8], values β=16, γ=4 have been first used for the categorization of low quality images. These parameters indeed greatly depend on the training corpus and different settings of their values produce a significant variation in performances.

Notice that, in Equation (1), features with negative difference between positive and negative relevance are set to 0. This is an elegant feature selection method: the 0-valued features are irrelevant in the similarity estimation. As a result, the remaining features are optimally used, i.e. only for classes for which they are selective. In this way, the minimal set of truly irrelevant features (giving 0 values for all the classes) can be better captured and removed.

In [3] a modified Rocchio model is presented that makes use of a single parameter γ_i as follows:

$$\Omega_f^i = \max\left\{0, \frac{1}{|R_i|}\sum_{d_h \in R_i} \omega_f^h - \frac{\gamma_i}{|\overline{R}_i|}\sum_{d_h \in \overline{R}_i} \omega_f^h\right\} \quad (2)$$

Moreover, a practical method for estimating the suitable values of the γ_i vector has been introduced. Each category in fact has its own set of relevant and irrelevant features and Eq. (2) depends for each class i on γ_i. Now if we assume the optimal values of these parameters can be obtained by estimating their impact on the classification performance, nothing prevents us from deriving this estimation independently for each class i. This result in a vector of γ_i each one optimising the performance of the classifier over the i-th class. The estimation of the γ_i is carried out by a typical cross-validation process. Two data set are used: the training set (about 70% of the annotated data) and a validation set (about 30% of the remaining data). First the categorizer is trained on the training set, where feature weights (ω_f^d) are estimated. Then profile vectors Ω_f^i for the different classes are built by setting the parameters γ_i to those values optimising accuracy on the validation set. The resulting categorizer is then tested on separated test sets. Results on the Reuters benchmark are about 85%, close to state-of-art more complex classification models ([3]). In Section "*Performance Evaluation*" the results as measured on the transcribed RAI news will be discussed.

5 Extending Internal Metadata with Web Material

5.1 Collecting External Evidence from the Web: The N_E Set

In RitroveRAI the task of creating a N_E set is achieved by mining several Web sites of news providers. Indexed web pages provide external news sets for each provider N_E^1, N_E^2, …, N_E^n. Finally, the $N_E = N_E^1 \cup N_E^2 \cup ... \cup N_E^n$ is the union of individual sets.

As a case of study, we considered only the Web site of the Italian newspaper "*La Repubblica*"[7]. It publishes news categorized by legacy metadata (e.g. a set of 8 newspaper categories). Moreover, it is refreshed as new articles are available with news items published in standard HTML and tagged with date information, e.g. "(January, 12, /2003)".

A simple spidering process has been developed, based on the Google's API, to retrieve all the documents published in a date. A temporal window is used. It is centered in the day of publication of the internal news and its symmetrical width is proportional to a parametric time span[8]. Accordingly, the temporal distance between the retrieved news and the source transcribed segment is considered as an inversely proportional ranking score. The main criteria is still the IR relevance score extended as follows.

The task of creating a link between elements in N_I and Web elements in N_E require to assess topical relevance and temporal proximity. Notice that the search vector extracted for internal news is used as a query for retrieval among the N_E set.

It is to be noticed that the IR engine is first run over a superset of the target Web pages in order to get general and reliable statistics for feature weighting (i.e. occ_n scores). Then the ranking function is modeled according to the following properties:

- Document similarity must be maximized
- The time distance D between the source segments and the Web page should be minimal
- The (RAI) topical category C^R of the segments should be coherent with the Repubblica category C^W for the Web news item

A comprehensive scoring model is as follows:

$$S(s,w) = sim(s,w) \cdot \frac{coh(C^R, C^W)}{D+1} \qquad s \in N_I, w \in N_E \tag{3}$$

where *sim* is the relevance produced by Lucene, and *coh* is a static function (*coherence table*) that measures the topical similarity between different categories (RAI, C^R, and "*La Repubblica*", C^W, respectively). The role of D promotes the alignment of "facts" happened in the same days. The category similarity *coh* refines the score and it is 1 only if the categories are the same, and lower as long as they tend to diverge: for example, it is almost 0 between RAI "*Sport*" and "*La Repubblica*" "*Foreign Affair*".

[7] `http://www.repubblica.it`

[8] Currently a time span of 2 days are used. Web pages outside such [–2,2] range are not considered for alignment.

Given a segment *s*, the alignment with a Web candidate *w* is finally accepted whenever the score *S(s,w)* is above a given acceptance threshold that has to be determined experimentally (see Section "*Performance Evaluation*").

5.2 Selecting Named Entities from Web Material: The NE_E Set

The aligned news provide evidence external to a segment able to trigger the extraction of more correct named entities as additional metadata.

The major problem in this phase is that the two source texts have quite different extensions. Usually Web data are excerpts of texts published on a newspaper and are longer: they discuss a "fact" in a broader way. This "additional" information is a suitable enrichment if:

- Misspellings in the transcriptions prevented the correct recognition of named entities that are alternatively found in the Web text;
- Time constraints in the TV news led to the exclusion of some relevant information (e.g. the name of a person entering into the underlying fact but not mentioned).

On the contrary, several aspects lead to consider carefully the external information:

- Newspaper articles discusses facts and opinions in a lengthy fashion, so that other facts and participants can be mentioned even when they are not directly related to the transcribed segment.
- The Web news are not perfectly aligned or time distance *D* is not 0: in this case new found named entities may be misleading.

The above observation lead us to apply filtering criteria to the acceptance of external Named Entities. A simple heuristic, based on a word distance metrics, has been developed.

An external Named Entity ne_e can be accepted if one of the following occurs:

- ne_e is also contained into the transcribed segment, i.e. $ne_e \in NE_I$;
- ne_e is repeated more than *m* times within the external Web news (appears to be central in the fact discussed in the aligned Web news);
- The named entity ne_e is close *enough* to other named entities that are also internal named entities, i.e. it exists one or more $n \in NE_I$ such that in the Web document

$$word\_dist(ne_e, n) < v$$

where *v* is a positive threshold. In this case the fact involving external and internal named entities is the same.

6 Publishing and Searching Enriched News

The publishing modules is responsible for showing the process results to the users, presenting them with a personalized profiles. The user logs in into the system identifying himself and implicitly declaring a filtering profile. This profile defines the categories of interests for the user.

The browser interface is shown in the following picture:

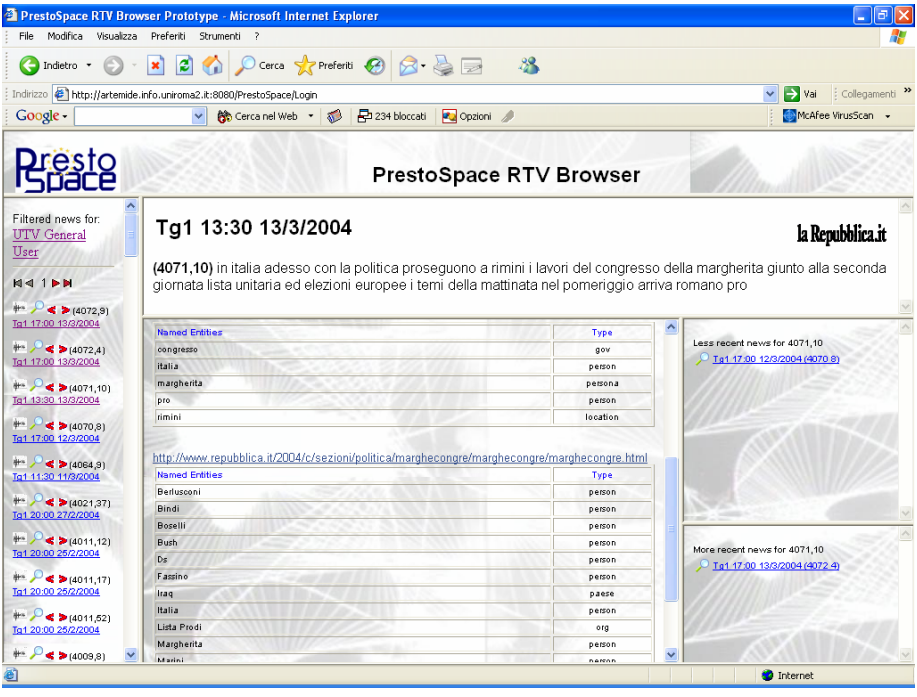

Fig. 3. A snapshot of the RitroveRAI browser

The left frame presents the entire broadcast news collection, with the ability to access to the audio recording. The user can access to the news information selecting its hyperlink. All the data associated to the news are shown in the central frame.

In the upper part of the central frame the transcribed segment is presented. The upper link refers to the best "*La Repubblica*" Web text aligned. The same link is shown in the middle of the bottom central frame (as a full URL).

Categorization is also reported in the bottom central frame as list of categories with their corresponding confidence factors (not visible in Fig. 3). The "legacy metadata" follow the categorization information, i.e. reporting data like the broadcasted segment, the day of recording, the TV channel or the tape IDs.

Finally, semantic metadata are presented (Fig 3, bottom central frame): one table for the internal metadata and one for the external ones. As we discussed before, the internal metadata are extracted using NLP over the broadcast news, and the external are collected using NLP over the aligned web news. These metadata are used to propose links to "related" broadcasted segments: in the right upper and bottom frames links are available to the set of less recent and more recent broadcasted news, respectively. These links are computed over the metadata and are built at run time according to user profiles.

7 Performance Evaluation: The RAI TV News

The RitroveRAI system has been tested on a large set of transcriptions from about one year of TV news (July 2003-June 2004). Segments of news have been used as source

information units for metadata extraction. The corpus includes about 20,676 of such units. The validation has been carried out for two independent tasks: news categorisation and Web alignment. Performances in the two cases have been estimated by the standard measures of *recall* and *precision*. *F-measure* as the harmonic mean of the two values as well as the *Break-even Point* (*BEP*), as the value for which recall is equal to precision, are also reported.

7.1 Evaluation of the RitroveRAI Text Categorization

The set of manually categorized news (annotated by RAI archivist) includes 1,861 segments. A split of 80% for training and 20% for testing has been imposed by random sampling the data set and balancing the different 26 RAI categories. Categories range from news specific classes (like "*Economics*" and "*Foreign Politics"*) to more general area like "*Health*". Each news was assigned to one or more classes, so that 2,328 assignments were available with an average rate of 1.25 class per news item. News items are distributed evenly among categories, so that only 11 categories had more than 80 members, an amount sufficient for reliable training. Validation has thus been carried out in two fashions: first by measuring the selection of the system among all the 26 classes, and then by restricting the testing to only the 11 reliable classes. Results (as *BEP* points) are reported in Table 1 for the 11 reliable classes.

Performance of the categorizer are good considering that a small subset of the archived material from RAI has been used for training. In particular small data sets penalize the categories that are more general (i.e. "*Employment/Job*") although more specific classes require less information to scale up to reasonable performances (e.g. "*Sport*", "*Life and Religion*"). When enough material is available the performances confirm the results of benchmarking (e.g. "*Politics*"). Notice how these measures are only based on tokens (bag-of-word modeling) of the transcribed news and how this material includes a significant amount of noise. Moreover, real-time categorization is ensured by the Rocchio model that, compared to more sophisticated text categorization techniques (e.g. Support Vector Machines), is much more efficient[9]

Table 1. Results (*BEP*) of the RitroveRAI Text Categorizer

Category	Training Set Size	BEP (26)	BEP (11)
Sport	76	0,83	0,72
Environment	55	0,45	0,56
Life and Religion	59	0,89	0,79
Current Events	172	0,45	0,54
Economics	149	0,60	0,76
Transportation	48	0,68	0,67
Foreign Affairs	518	0,75	0,78
Justice	346	0,61	0,67
Employment/Job	62	0,55	0,52
Politics	437	0,80	0,79
Health	58	0,73	0,46

[9] Profile based classification requires a number of scalar products tight to the number of classes that is much lower to the number of documents.

7.2 Evaluation of the RitroveRAI News Alignment

The validation of the Web alignment capability of RitroveRAI has been carried out on a reference set of about 410 news items (i.e. segments in transcriptions) manually annotated. The annotation has been added by a team of three archivists with a judgment about each of the candidate alignment in four classes: "*bad*", "*fair*", "*good*" and "*very good*". "*very good*" expresses an exact correspondence between the event/fact described in the two documents. As the focus of the Web material can be slightly different from the TV news, degrading levels of evaluation express overlaps of decreasing size: "*good*" is a valid correspondence but between a shorter transcribed news than the longer Web document with many more facts. "*fair*" reflects the same specific topic (e.g. "*Iraki war*") but possibly not the same fact. "*bad*" refers to clear mistakes of the links. In order to study the accuracy of the thresholds imposed to Eq. (3) annotators were presented with all the links receiving a score greater than 0[10].

In the evaluation we wanted to focus on news transcriptions of reasonable quality, i.e. significant segments to accurately measure the linking accuracy. We distinguish between "*monothematic*" and "*multithematic*" units, i.e. segments reporting just one or many more facts, respectively. Multithematic segments are usually due to wrong segmentations that groups two or more facts. Annotators found 308 monothematic and 102 multithematic segments. Data reported will refer only to the 1,587 alignments proposed for the monothematic segments.

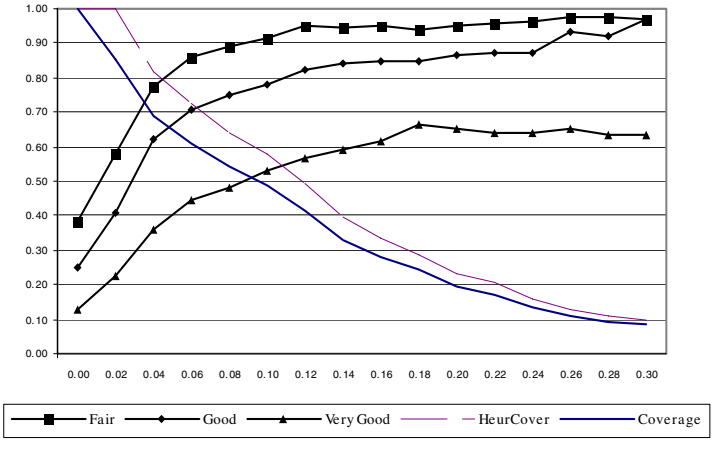

Fig. 4. Precision and Coverage of the RitroveRAI hyperlinking

Two performance indexes have been used: *precision* at the three levels of evaluation and *coverage*. Precision is the ratio between the number of links that received an evaluation equal or better then the level (from "fair" to "very good") and the total number of links proposed by the system. Figure 4 plots the three measures

[10] Notice that recall here does not apply: the annotators did not analyse the full "La Repubblica" Web site in the target time windows so that the gold standard set of all Web news valid for the alignment is not available.

according to the thresholds of acceptance imposed to the IR alignment scores (Eq. (3)). As the trends of all the curves suggest, there is a strong correlation between the thresholds and the accuracy. As a contrastive measure we computed the *coverage* as the ratio between the number of segments receiving at least one link and the total number of monothematic segments (i.e. 308). In Fig. 4 we see that coverage decreases smoothly and a kind of breakeven point is reached in the range of 65-75% precision. This is a quite good result if compared with standard performance of IR systems. Of course the constraints imposed on the alignment (in particular, the dates) are quite effective. Moreover, it must be said that not all the segments can be aligned by the system as (1) they may be not present on the "*La Repubblica*" Web site or (2) segments can be too short for significantly express a full fact.

An analysis of the optimal threshold (around 0,03) has been thus carried out. By imposing such a threshold we found that the amount of news not receiving any link is 62 (about 20%). However, among these segments we found 48 segments that are receiving only "bad" links. An analysis of about 20 such segments revealed in fact that there were no Web pages suitable for the alignment on that date (in the adopted source, i.e. "*La Repubblica*"). TV news may be in fact on local or curiosity information and some of them are not even mentioned on newspapers. It is likely that all the 48 segments were not to be linked at all. Correctly, a default threshold of 0.03 would have been prevented all the erroneous links to be proposed. Accordingly, we removed those 48 segments from the testing data set (i.e. the 308 monothematic segments) obtaining a reduced set of segments (302-48=254). This simulates a system with an heuristic threshold that correctly assigns no link to the above 48 candidates. Evaluation of such a system would be focused only on the 254 test segments with an alternative coverage plot ("Heuristic coverage" in Fig. 4) that is slightly higher than the previous. Notice how the precision plots for such a modified system (by imposing every threshold 0.03 or higher) do not change for any acceptance rate.

8 Conclusion and Future Work

In this paper the RitroveRAI system for the extraction of semantic metadata from broadcasted TV and radio news has been presented. Human Language Technologies are here exploited to extract from the news transcriptions grammatical and semantic information and align them with Web documents. Alignment with these latter well-formed texts is used to validate the extracted metadata as well as to complete them with additional information.

The result is a metadata repository that supports querying in plain natural language (e.g. "*Bush commenting Irak elections*") as well as more conceptually motivated languages (e.g. comment(*Bush*:`agent`,*'Irak elections'*:`theme`)). In the originating European projects, PrestoSpace, work is in progress to integrate the RitroveRAI language processing functionalties with ontology services, like Named Entity classification and semantic-driven coreference resolution in texts, made available by other partners (the KIM ontology, [11]). As the involved NLP technologies (in particular, the CHAOS parser) support text processing in two languages, Italian and English, the RitroveRAI system already enables extraction of language-neutral semantic metadata and multilingual information access: querying in English can be parsed and normalized by CHAOS so that metadata can be searched in a language neutral manner for cross-lingual reitrieval.

Experimental work presented in this paper has been carried out on a significant scale (hundreds of segments and thousands of links) and demonstrates viability for large scale processing. The client-server Web architecture of the system is currently under testing to process TV broadcasted by RAI, daily. Categorization, although trained over a limited test set, is currently running with an acceptable accuracy. More importantly, the Web alignment method proposed reaches high level accuracy. This opens more space for the extraction of deeper phenomena from Web, like event descriptions with recognition and role assignment to participant of the detected events.

Open problems refer to the improvements needed to deal with noisy input, i.e. wrong news segmentation. All the current RitroveRAI processing is based on the strict assumption that segmentation is provided as a form of preprocessing. After alignment however, more semantic information is available to the system for some analysis of odd segments (e.g. too short or too long): algorithms for a posteriori merging and splitting can thus be made available. This task is close to automatic segmentation of long documents as carried out in text summarization ([1], [10] or [12]). In particular Lexical Chains ([1]) and Latent Semantic Analysis ([8],[6]) can be here applied either to the TV segments or to their alignments on the Web. Integration of these two (independent) information sources will capitalize further the alignment to improve segment detection as well as all the subsequent processing steps.

The browsing capabilities of the RitroveRAI system are already supporting natural language querying and user specific Web browsing (as in Fig. 3). Moreover, as mentioned in the introduction, significant portions of the system are adaptive, including categorization and Web alignment. This makes RitroveRAI a typical example of large scale adaptive Semantic Web application. Its capabilities for IE and automatic Web alignment coupled with its browsing and querying modalities are a feasibility proof of a new generation of multimedia information brokering systems over the Web.

Acknowledgement

The authors want to thank RAI, Centro Ricerche ed Innovazione Tecnologica (CRIT) of Torino (Italy), and in particular the staff involved in PrestoSpace, Giorgio Dimino, Daniele Airola Gnota and Laurent Boch, for having made available the data set for training and testing and for the helpful support to the architectural and application design choices.

References

[1] R. Barzilay, M. Elhadad, *Using Lexical Chains for Text Summarization*. In the Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, 1997.

[2] Basili, Roberto, Pazienza, Maria Teresa, Zanzotto, Fabio Massimo, *Efficient Parsing for Information Extraction*, Proceedings of the European Conference on Artificial Intelligence (ECAI98), Brighton, UK, 1998.

[3] R. Basili, A. Moschitti, M.T. Pazienza, "*NLP-driven IR: Evaluating performance over a text classification task*", In Proceeding of the 10th "International Joint Conference of Artificial Intelligence" (IJCAI 2001), August 4th, Seattle, Washington, USA 2001.
[4] Basili R., F.M. Zanzotto, Parsing Engineering and Empirical Robustness, 8 (2/3) 97120, Journal of Language Engineering, Cambridge University Press, 2002
[5] F.Y.Y. Choi, P. Wiemer-Hastings and J. Moore. "Latent semantic analysis for text segmentation". In Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing, pp. 109- 117, 2001
[6] Vasileios Hatzivassiloglou, Judith Klavans, and Eleazar Eskin. 1999. *Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning.*
[7] Otis Gospodnetic. 2003. Advanced Text Indexing with Lucene. http://lucene.apache.org
[8] David J. Ittner and Lewis, David D. and David D. Ahn, *Text categorization of low quality images*, Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, 1995, Las Vegas, US, 301—315.
[9] Landauer, T. K., Foltz, P. W., & Laham, D., *Introduction to Latent Semantic Analysis*. Discourse Processes, 25, 259-284, (1998).
[10] Daniel Marcu. 1999. *The automatic construction of large-scale corpora for summarization research*. In Proceedings of SIGIR 99.
[11] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, Angel Kirilov, Miroslav Goranov. *KIM – Semantic Annotation Platform*. 2nd International Semantic Web Conference (ISWC2003), Florida, USA, 2003.
[12] Hongyan Jing. 2002. *Using hidden Markov modeling to decompose human-written summaries*. Computational Linguistics, 28(4):527–543