Context-driven Semantic Enrichment of Italian News Archive*

Andrei Tamilin, Bernardo Magnini, Luciano Serafini, Christian Girardi, Mathew Joseph, and Roberto Zanoli

FBK, Center for Information Technology - IRST
Via Sommarive 18, 38050 Povo di Trento, Italy

Abstract. *Semantic enrichment* of textual data is the operation of linking mentions[1] with the entities they refer to, and the subsequent enrichment of such entities with the background knowledge about them available in one or more knowledge bases (or in the entire web). Information about the context in which a mention occurs, (e.g., information about the time, the topic, and the space, which the text is relative to) constitutes a critical resource for a correct semantic enrichment for two reasons. First, without context, mentions are "too little text" to unambiguously refer to a single entity. Second, knowledge about entities is also context dependent (e.g., speaking about political life of Illinois during 1996, Obama is a Senator, while since 2009, Obama is the US president). In this paper, we describe a concrete approach to *context-driven semantic enrichment*, built upon four core sub-tasks: detection of mentions in text (i.e., finding references to people, locations and organizations); determination of the context of discourses of the text, identification of the referred entities in the knowledge base, and enrichment of the entity with the knowledge relevant to the context. In such approach, context-driven semantic enrichment needs also to have contextualized background knowledge. To cope with this aspect, we propose a customization of Sesame, one of state-of-the-art knowledge repositories, to support representation and reasoning with contextualized knowledge. The approach has been fully implemented in a system, which has been practically deployed and applied to the textual archive of the local Italian newspaper "L'Adige", covering the decade of years from 1999 to 2009.

Keywords: knowledge extraction, context-driven semantic enrichment, contextualized knowledge representation

1 Introduction

The exploitation of background knowledge for text understanding is nowadays becoming a very appealing research area due to the wide availability of large

* The original publication is available at http://www.springer.com.

[1] The terms "mentions" and "entities" have been introduced within the ACE Program (Linguistic Data Consortium, 2004). "Mentions" are equivalent to "referring expressions", while "entities" are equivalent to "referents", as widely used in computational linguistics.

sources of structured background knowledge in form of semantic web data, as well as huge amount of unstructured textual data. Our intuition on how textual understanding using background knowledge should be implemented is by means of a two phase *loop* in which (i) knowledge is automatically extracted from text by exploiting some form of pre-existing background knowledge and (ii) the extracted knowledge is used to extend the background knowledge and in this way it will contribute to extract new knowledge from text. In this paper we focus on phase (i) and we show how the usage of the information about *context* can contribute to improve the quality of that phase.

Our approach is based on the observation that when humans read a piece of text, they exploit their capability of establishing a link between the mentions, which occur in the text, and knowledge they have previously acquired about corresponding entities mentions refer to (such as people, organization, locations, events, and etc.). The impossibility to create such a link prevents people from combining their previous knowledge with the information contained in the text. Why people can actually do this is by taking into account the context in which each mention occurs in. Information about the context (as for instance that the article is about a soccer match which took place in Milan) helps them to focus their attention, and to limit the scope of the knowledge they remember. For instance, if we are reading an article on the newspaper speaking about the certain soccer match between Milan and Juventus, which is part of the Italian League and was taken in Turin on the 20 of October 2000, reading the string "Boban", if we already know him, we will focus our attention on an ex-Yugoslavian soccer player Zvonimir Boban, who played with Milan from 1992 to 2001, considering the fact that he played the role of midfielder, and all the other knowledge about this person relevant given the context of the article. In this process, the context of discourse helps us to focus our attention on the soccer player and not Boban Marković, the Serbian trumpet player. Similarly, the context will limit the "activation of knowledge" to one necessary to understand the information contained in the article. So for instance, it will not be necessary to remember that before 1990 Boban played in Dinamo Zagabria team.

To implement the above schematic model into an automatic program, we should be able to realize the operation called *semantic enrichment* of textual data. This is the operation of automatically linking mentions with the entities they refer to and the subsequent enrichment of such entities with the background knowledge about them available in one or more knowledge bases (or in the entire web). Information about the context in which a mention occurs, (e.g., information about the time, the topic, and the space, which the text is relative to) constitutes a critical resource for a correct semantic enrichment for two reasons. First, without context, mentions are "too little text" to unambiguously refer to a single entity. Second, knowledge about entities is also context dependent (e.g., speaking about political life of Illinois during 1996, Obama is a Senator, while since 2009, Obama is the US president), and we need to know the context in order to identify the correct portion of knowledge the mention should be enriched with.

This paper defines and provides a concrete implementation of the *context-driven semantic enrichment* of text. Context-driven semantic enrichment is built upon four core sub-tasks:

- mention detection in the text (we will focus on people, locations and organizations);
- determination of the context of discourse of the text;
- identification of the referred entities in the knowledge base;
- enrichment of the entity with the knowledge relevant to the context.

Practically, this work describes the approach and implemented system for context-dependent semantic enrichment of Italian news archive. The approach integrates natural language processing tools for extraction of named entities with the background knowledge expressed in semantic web standards, such as RDF/OWL. In order to enable context-sensitive enrichment of the extracted named entities, we have extended the standard state of the art semantic repository Sesame for managing RDF with the capability of expressing contextually-qualified RDF/OWL knowledge bases, focusing on temporal, geographic and thematic contextual dimensions. Formally, for contextualization of background knowledge we adopted and extended the state of the art context as box representation framework [9], in which the contextual space is defined by a fixed number of partially-ordered contextual dimensions and the concrete context, containing RDF/OWL knowledge base, is defined by a vector of values the corresponding dimensions take. By virtue of dimension orders, in such a framework contexts automatically exhibit generalization/specialization ordering allowing further effectively localize contexts possibly containing knowledge relevant for enrichment of the given textual source.

The paper is further organized as follows. In Sect. 2 we describe the processing pipeline of the semantic enrichment process. Employed knowledge extraction techniques are described in Sect. 3. In Sect. 4 we present the approach to organize background knowledge in a context-sensitive way, shedding the light on the formal aspects of the approach. In-detail description of the context-dependent semantic enrichment process is depicted in Sect. 5. Practical issues related to application to real data set are described in Sect. 6. Relation with other similar approaches and systems is discussed in Sect. 7. Finally, Sect. 8 concludes and outlines the future work.

2 Enrichment Pipeline by Example

Semantic enrichment of a text aims at making available background knowledge that is supposed to be relevant while reading and interpreting that text. As an example of context-driven semantic enrichment, suppose we read the following piece of text about a soccer match:

```
Milan - Juventus (Friday, November 20, 2000)
```

"Milan was unlucky to hit the post with a Boban header in the first half but came out of the dressing room determined to score and win all three points."

In this case relevant background information about "Boban" would include knowledge, among the other, about his team (i.e., Milan) at that time (i.e., 2000), his role in the team (i.e., midfielder). We might also be interested in background knowledge about the specific match (i.e., Milan-Juventus), or about the town in which the match has been played (i.e., Milan).

Now, suppose we encounter "Boban" in the context of a different match, like the one reported in the following text.

PARIS (Thursday, July 9, 1998)

"Croatia captain Zvonimir Boban won the ball at the top of the box at Croatia's defensive end. Thuram, a right back who had made a long run forward, didn't give up on the play, coming behind the dawdling Boban and knocking the ball loose."

In this case, different background knowledge about the same person (i.e., Boban) should be selected. Specifically, the team, at that time (i.e., 1998), would be Croatia, the match is France-Croatia, and the town of the match is Paris.

In order to provide appropriate background knowledge, the following steps are necessary: (i) some understanding of the entities mentioned in the context, including the fact that "Boban" is the name of a person and that "Milan" is the name of an organization: this is referred in the literature as *named entities recognition*; (ii) understanding that "Boban" in the first context and "Zvonimir Boban" in the second one, actually denote the same person: this is referred as *cross-document coreference*; (iii) being able to recognize that the person mentioned in the two contexts is included in a repository of background knowledge: this is referred as *entity disambiguation*, because there can be more than one person with the same name; (iv) being able to select the correct portion of knowledge available in the background repository with respect to the actual textual context, for instance the composition of the Milan team in the first example and the composition of the Croatian team in the second: this is called *context selection*, and (v) being able to attach correct pieces of knowledge to corresponding textual portions: this is referred as actual *enrichment*.

3 Automated Extraction of Named Entities

The ability to recognize and classify Named Entities (Named Entity Recognition), such as people and locations names, is an important task in various areas, including topic detection and information retrieval. Cross-document coreference extends the task into deciding whether or not different mentions refer to the same entity, and the task becomes more complex when documents come from different sources, probably having different authors, conventions and style. Note

that an entity (such as Valentino Rossi, the MotoGP World Champion) can be referred to by multiple surface forms (e.g., Valentino Rossi, Rossi and Valentino) and a surface form (e.g., Rossi) can refer to multiple entities (e.g., Vasco Rossi, the Italian rock star, and Paolo Rossi, the famous ex-football player); performing this task allows users to get information about a specific entity from multiple text sources at the same time.

3.1 Named Entity Recognition

Named Entity Recognition is a subtask of Information Extraction which aims to classify words in text into predefined categories. Examples of named entities are person names (PER), location (LOC) and organization names (ORG). Spurred on by the Message Understanding Conferences (MUC), a considerable amount of work has been done in last years on the Named Entity Recognition and a number of machine-learning approaches were proposed, such as: Hidden Markov Model (HMM), Support Vector Machines (SVMs), and Conditional Random Fields (CRFs). Drawing from our participation at Evalita 2007[2] and at ACE08[3], we built Typhoon [13], a system for Named Entity Recognition in which two different classifiers based on CRFs and HMM are combined in cascade to exploit global features such as *Data Redundancy* and *Patterns* extracted from a large text corpus of about one billion of words. *Data Redundancy* is attained when the same entity occurs in different places in documents, whereas *Patterns* are 2-grams, 3-grams, 4-grams and 5-grams preceding, and following recognized entities in the large corpus. The system can use additional features, such as that given by a Text Classifier able to recognize the category to which the story belongs (e.g. sport, economy). Typhoon consists of two classifiers in cascade, but it is possible to use a single classifier making the system faster (100 times faster, with a speed rate of about 20,000 tokens/sec); whereas the second classifier will be used in combination to the first one when more accuracy is needed. The system took part in Evalita 2009 Named Entity Recognition task [10] for Italian language performing as the best tagger (see Table 1).

Table 1. Evalita 2009 results

Entity type	Precision	Recall	F_1
PER	90.29	86.42	88.31
LOC	86.12	84.16	85.13
ORG	71.71	69.43	70.56

[2] http://evalita.fbk.eu/2007/
[3] http://www.itl.nist.gov/iad/mig/tests/ace/

3.2 Cross-document Coreference

To corefer the Named Entities we used the system developed by Popescu and Magnini [12] in cascade to Typhoon. The system is based on agglomerative clustering technique able to exploit other Named Entities and professional categories co-occurring with the ambiguous entity in the same document; it was first tested in the SEMEVAL 2007 Web People Search task, performing the second best result among 16 systems with the following performance in terms of the harmonic mean of purity and inverse purity, which are standard clustering evaluation metrics: F_1=0.77 (Purity=0.75 and Inverse Purity=0.80).

4 Contextualized Background Knowledge Repository

The recognition of the fact that most of the knowledge available on the semantic web in form of RDF/OWL data is tailored to a specific context of use, has fostered recently the investigation on practical extensions of the semantic web languages to make explicit the representation of context associated to a knowledge resource. The use of explicit qualification of knowledge with contextual information has been investigated for such issues as provenance and trust of data [7, 4], expressing propositional attitudes [11], dealing with temporally-stamped data [8], and access control [5]. In this work, we propose a lightweight *contextualized background knowledge repository* for managing and querying RDF/OWL data qualified with a set of contextual restrictions, practically limiting the scope to the temporal, geographic and thematic boundaries since those can be directly derived from a textual material.

4.1 Representation Framework

For representation of contextually qualified RDF/OWL knowledge we formally adopted and extended the state of the art *context as a box* framework [9]. According to this framework, a *context* is defined as a set of logical statements, or a knowledge base, inside the box, and an array of *contextual dimensions*, outside of the box. For example, if $\mathcal{C}$ is the context of the current Italian parliament, it can contain the information $\mathsf{primeminister}(\mathsf{berlusconi})$ and the parameters are for instance $\mathsf{time}(\mathcal{C}) = \mathsf{8may2008–now}$, $\mathsf{location}(\mathcal{C}) = \mathsf{Italy}$, $\mathsf{subject}(\mathcal{C}) = \mathsf{Politics}$.

In the present work, we pursue the additional requirement to the context as box framework demanding the values of each of contextual dimensions to be taken from structured domains with defined on them broad-narrow relations. For instance, the values for $\mathsf{time}(\mathcal{C})$ are time intervals, the values of $\mathsf{location}(\mathcal{C})$ are geographical regions, and the values of $\mathsf{subject}(\mathcal{C})$ are topics. For time and location dimensions the broad-narrow relation can be naturally defined as the interval and region containment respectively, while for subject dimension the topic-sub-topic relation can be considered.

As an example, Fig. 1 depicts a contextualized repository composed of three contexts describing governments in Italy. It can be easily observed that the context $\mathcal{C}_1$ is broader than contexts $\mathcal{C}_2$ and $\mathcal{C}_3$.

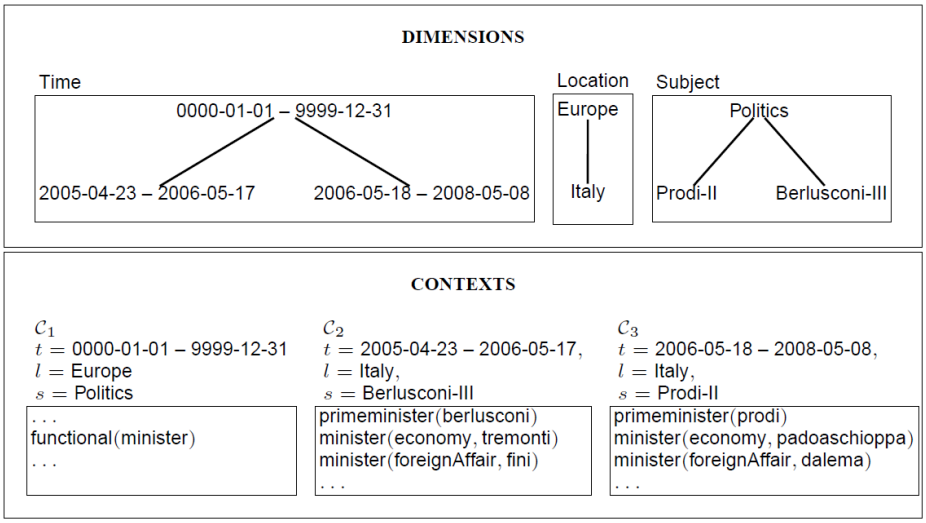

Fig. 1. Example of contextualized repository

4.2 Querying for Contextualized Knowledge

In answering queries, knowing the scope of the query is quite crucial. For instance, if one asks for the prime minister in the context of 2005 United States Politics or in the context 2005 Italian Politics he clearly obtains two different answers. From this elementary example one can see how relevant the context of the query for providing the right answer. That is why a query to a contextualized knowledge repository is in fact a contextualized query, composed from the query itself and also from the (set of) context(s) in which the query should be evaluated.

The choice of the most appropriate context for executing contextualized query is very crucial factor. Sending a query to the "wrong" context can produce an empty answer. On the other hand to precisely determine the correct context at which a query should be sent is in most of the cases a difficult task. To mitigate this problem, we introduce the notion of a *query shifting*, which is the operation of redirecting a query from one context to another relevant context, when query fails to find any fruitful information in the current context. The hierarchical structure of contexts, induced by partial orders of the context dimensions, provides the basic graph on which queries are shifted across contexts. More specifically, as a semantic metric of context closeness the relation "directly covers" and "is directly covered by" are exploited.

Let us see an example of such a shifting considering the contextualized knowledge repository shown in Fig. 1. Suppose we want to submit to the repository a query for knowing who was the minister of economy during the period 2005–2009 in Italy. This request can be encoded into the contextual query $(\mathsf{minister}(\mathsf{economy}, x), \langle \mathsf{2005–2007}, \mathsf{Italy}, \mathsf{Politics} \rangle)$. Since the context with dimen-

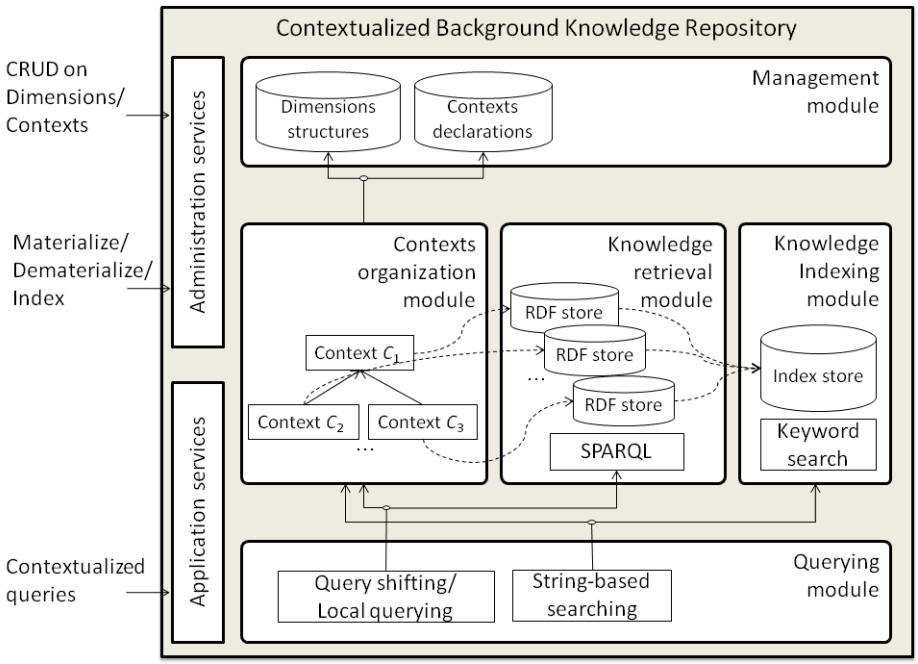

Fig. 2. System architecture of the contextualized repository

sions $\langle$2005–2007, Italy, Politics$\rangle$ does not exist in the repository, then a new context $\mathcal{C}_4$ is created with empty content and its position in the context graph is computed (namely $\mathcal{C}_4$ position is below $\mathcal{C}_1$ and above $\mathcal{C}_2$ and $\mathcal{C}_3$). The query in $\mathcal{C}_4$ will not produce any result. On the other hand if, we "zoom" into the contexts directly covered by $\mathcal{C}_4$, namely $\mathcal{C}_2$ and $\mathcal{C}_3$, and submit the same query to both of them, we will obtain the following contextualized answers:

Context	Answer
2006-05-18 – 2008-05-08, Italy, Prodi-II	padoaschioppa
2005-04-23 – 2006-05-17, Italy, Berlusconi-III	tremonti

4.3 System Architecture

Implementation architecture of the contextualized knowledge repository is graphically depicted in Fig. 2. The repository is composed of four principal modules, whose functionalities are briefly described below.

The management module supports the functionalities for (a) defining and managing the dimensions structures and (b) defining and managing the set of contexts comprising the repository.

Contexts organization module exploits dimensions structures in order to compute context covering relation for the organization of contexts of the

repository. Practically, context cover relation has been represented by a Hasse diagram, one of the popular representations of partially ordered sets. This structure is used in order to compute the context pairs among which it is possible to shift queries.

Knowledge retrieval module performs the actual loading of knowledge of contexts into the storage, i.e., materializing knowledge, for further execution of knowledge retrieval queries.

Knowledge indexing module performs the textual indexing of the knowledge contained in the materialized repositories to enable string-base search queries.

Querying module enables to answer contextualized queries using SPARQL and keyword-based search.

Practically, we grounded our prototype on Sesame RDF storage and querying framework [3], which is one of the most popular free open-source tool having good performance and stability.[4] For indexing and text-based searching we used open-source Apache Lucene Solr platform.[5]

5 Context-driven Semantic Enrichment

Having on one side the textual material, annotated by automated entity extractor with textual mentions (referring to people, locations and organizations), and the populated contextualized background knowledge repository, the task of a context-driven semantic enrichment consists in mapping mentions to the entities in the repository and then retrieving the knowledge about these entities relevant to the textual context mentions occur in. In the following we discuss major steps in more details.

5.1 Entity Disambiguation

The first step to entity disambiguation consists in identification of the name for the entity from the multiple possible ways it is referred to in the text. For this task from the computed coreference clusters we select the longest and the most frequent mention occurring in the text. This simple heuristic in practice allows bringing from the text the most representative, complete name for the frequently occurring entities, which is the combination of name and surname for people, non abbreviated name for locations and organizations. Using the complete name allows to establish the accurate match to the entities available in the knowledge repository. Practically, the matching is performed by querying the textual index constructed from ontologies of the repository using identifiers and available textual labels attached to ontological individuals. However, in the general setting, due to the name ambiguity problem, this procedure brings from

[4] In the future we plan to investigate the use of OWLIM extension to Sesame for OWL data; more details on OWLIM can be found at http://www.ontotext.com/owlim

[5] http://lucene.apache.org/solr/

the repository the set of entities having the same name but denoting different real-world entities. To eliminate further the entity ambiguity we employ the analysis of the text, the entity mentions occur in, in order to guess the imposed contextual constraints.

5.2 Context Selection

Topic and time are constraints we practically employed for identification of the entity context from text. Selection of topic is based on the use of automatically extracted from text keywords (using the algorithm presented in [1]) allowing to fall the textual source into the fixed list of categories, ranging from broad categories, such as for instance sport and politics, to the more detailed ones, such as soccer, soccer series and annual championships, government legislatures, and etc. For the identification of temporal constraint we at the moment adopted the simple strategy of using the publication year of the text. Currently we are working on automatic detection of temporal expressions from text for more accurate determination of the temporal constraint.

Identified contextual constraints are used for further filtering the list of matching entities and consequent retrieval of the relevant background knowledge from the corresponding contexts of the contextualized repository.

5.3 Enrichment

Using the detected entities and guessed contextual constraints, the concluding semantic enrichment phase extracts from the corresponding contexts of the repository the background knowledge on the entity. This later practically evaluates to execution of the SPARQL query asking for triples having in subject or object the given entity.

6 Practical Deployment

The presented context-driven semantic enrichment approach has been fully implemented in a system, which has been practically deployed and applied to the textual archive of the local Italian newspaper "L'Adige"[6]. The archive contains 620,641 articles in Italian from January 1st 1999 to October 15th 2009 ranging over the regional and national news in the domain of politics, sports, economics, culture, education, and etc. Practically, for the processed archive we have implemented the portal giving the possibility to search the news archive by entities addressed in the articles, see the relevant background knowledge attached to entities by semantic enrichment when reading the articles, as well as see the summary card assuming all background knowledge on interested entity. The screen shot of the implemented portal is shown in Fig. 3. In the following we give some figures describing the steps of the enrichment process.

[6] http://www.ladige.it

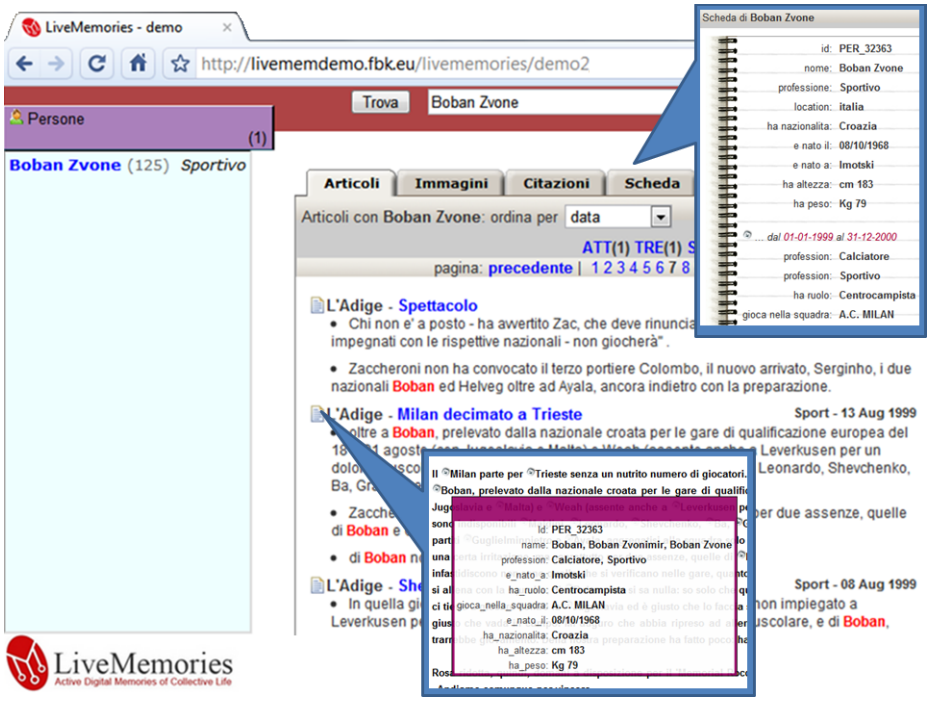

Fig. 3. Semantically enriched news portal for "L'Adige" archive

6.1 Extraction of Named Entities

For automated extraction of named entities from "L'Adige" dataset we practically used the Typhoon system described in Sect. 3. From total amount of 257,255,240 tokens constituting the dataset we have extracted 5,169,188 mentions referring to people; 2,977,960 - to organizations and 2,958,228 - to locations. We consider the reported in Sect. 3 evaluation obtained at Evalita a reasonable indicator of the accuracy of the extracted entities; considering also the fact that at Evalita systems were evaluated on a subset of documents extracted from "L'Adige" archive.

From the extracted Named Entities we further constructed 133,906 clusters corresponding to distinct persons, 34,270 clusters of organizations and 9,836 - of locations using the integrated in Typhoon coreference resolution algorithm by Popescu and Magnini [12]. The coreference system limited to deal with entities of type PER only has been extended by us to work also with ORG and LOC too. The system accuracy was computed on a gold standard constructed on a subset of "L'Adige" news stories from 1999 to 2006, containing 209 person names corresponding to 709 entities, for a total of 43,704 annotated documents [6]. On this corpus the system performed as F_1=0.80 (Purity=0.91 and Inverse Purity=0.75).

6.2 Feeding the Background Knowledge Repository

Looking for semantic resources on Italy and in Italian language we recognized the very limited availability of such models in RDF/OWL on the web. To bootstrap the construction of the background knowledge in Italian, we have developed a set of procedures for semi-automatic conversion of existing semi-structured resources. In particular, we used various public classifications of National Statistics Institute (ISTAT)[7], data tables of national and provincial Economical Registries (Camera di Commercio)[8], official web sites of Italian senate and chamber[9], National Olympic Committee (CONI)[10], and etc.) into the formal ontologies expressed in RDF/OWL. The ontologies has been constructed in Italian language and focused on the following domains of interest: sport (players, teams, regular championships and competitions in soccer, auto- and motosport, ice hockey, and others), education (Italian educational system and degrees, universities, and others), economy (economical activities and professions, industries and craftsmen, banks, and others), politics (Italian political system, national and regional legislatures, deputies, senators, political parties, and others), and geography (administrative division of Italy into regions, provinces, communes, detailed geographic composition of the Province of Trento in accordance with the Toponymy database[11]). All of the ontologies have been loaded into proper contexts of the repository presented in Sect. 4.

6.3 Semantic Enrichment of Named Entities

With relatively simple effort spent for creating the collection of knowledge for the contextualized repository, the execution of the semantic enrichment over the complete "L'Adige" dataset 45% of news articles has been enriched (272,253 article from total of 600,470 distinct news articles), counting if at least one recognized in article entity of type PER, LOC or ORG has been enriched. Clusters produced by coreference resolution algorithm have been mapped to the corresponding entities in the repository with the following numbers: 35% of people clusters, 42% of location clusters, and 2% of organization clusters. In order to perform the qualitative assessment of the performed enrichment we are planning to create in the near future the corresponding gold standard. Preliminary, using the gold standard for cross-document coreference evaluation of people [6], we have first manually attached to famous people clusters (i.e., those with ambiguous names, such as Paolo Rossi, Roberto Mancini, and etc.) their corresponding person entity available in the current version of the repository; and then we have performed the enrichment of the gold standard articles. As a result we got a fairly high precision of 91% when enriching these famous people due to the ability to recognize the specific detailed context (such as championship of the

[7] http://www.istat.it/dati/
[8] http://www.cameradicommercio.it/
[9] http://www.senato.it/, http://www.camera.it/
[10] http://www.coni.it/
[11] http://www.trentinocultura.net/territorio/toponomastica/

certain year in sport, or certain political legislature). Straightforwardly, when we disabled the consideration of context and performed the enrichment against the whole knowledge repository the precision fall down, because of the big number of wrong matches produced due to the name ambiguity problem.

7 Related Work

During the last decade the introduced semantic web vision has fostered the development and public sharing on the web of numerous structured resources containing the background knowledge on different domains of interest. Availability of high quality knowledge consequently drew attention to the development of systems for exploiting this knowledge. The task of semantic enrichment of text is one of such applications aiming at making computers comprehend and integrate the knowledge contained in text. From the number of integrated solutions for the semantic enrichment reported in the literature, we would like to mention several remarkable approaches which implement the complete semantic enrichment pipeline and are close to the one presented in this work, namely exploiting the notion of context.

In [2] the authors present the approach to using DBPedia for enriching and interconnecting the number of data sources within BBC.[12] The main idea of their system is to analyze the web pages on music offerings, TV channels and programs in order to provide contextual, semantic links for connecting and navigating the content described in different pages using the entities corresponding to artists, bands, musical albums, concerts, and etc. The interlinking approach with DBPedia exploits the notion of context' for disambiguation; the idea is to group entities co-occurring with one another in text and further to search for corresponding entities in the DBPedia such that they all fall into corresponding similarity cluster.

Enrycher, described in [14], is another similar example of the completely implemented pipeline for semantic enrichment of textual resources with the knowledge from DBPedia. In a similar vein, it exploits the notion of context in form of a group of co-occurring in text entities and further seeks to find a cluster of corresponding DBPedia entities matching those in context.

The main difference of the present study is that the background knowledge is "clustered" by construction of the repository into the contexts and the detection of context from a textual resource is based on its temporal and thematic analysis, rather than on co-occurrence of entities.

8 Conclusion and Future Extensions

In this work we presented the approach and implemented system for semantic enrichment of textual materials written in Italian with the background knowledge contextually relevant to a given text. A number of possible improvement steps

[12] http://bbc.co.uk

are planned for the future to increase the accuracy of the current version of the system and improve the overall utility and usability. On the one side, this means to improve the quality of extraction and enrichment, on the other side, this means to construct the practical services benefiting from the performed enrichment.

For entity disambiguation we currently considered only the mentions of type proper name available in text for finding matching entity in the knowledge repository. However, the state of the art named entity recognizers, as Typhoon we have used in the current implementation, are capable of extracting richer set of mentions, in particular, nominal expressions practically referring to the professions in case of people (such as "the mayor of Trento" for the entity "Alberto Pacher"), types of locations (such as city or lake), and types of organizations (such as gmbh or srl). We plan to elaborate that information in the future for establishing more accurate link to the repository.

The accuracy of the semantic enrichment crucially depends on the quality of the context detection from text: the topic, time and location. One of the simplifications in the proposed context detection scheme is the identification of relevant time by considering only the publication date of the text. We are working on attaching to the pipeline the recently developed Chronos system for automatic recognition in text of temporal expressions, and using them further to identify the temporal span of the article from these temporal expressions in a more sensitive, granular way.

As of the improvement of the background knowledge collection, we are currently exploiting the use of other publicly available structured resources of the background knowledge and the ability to import those into the presented contextualized knowledge repository. We are particularly interested in a high quality knowledge bases, such as for example Freebase[13] and DBPedia[14]. The crucial point for their applicability to the proposed semantic enrichment flow consists in the ability to partition these resources into the contexts, identified by temporal, spatial and thematic dimensions.

For the qualitative assessment of the enrichment procedure, in the near future we plan to work out the evaluation methodology and construct the gold standard for people/location/organization enrichment on top of the corresponding gold standard for evaluation of cross-document coreference.

Acknowledgements

This work has been supported by LiveMemories project (Active Digital Memories of Collective Life) funded for 2008-2011 by Autonomous Province of Trento under the call "Major Project". The authors express additional gratitude to Silvana Marianela Bernaola Biggio, Elena Cardillo, and to all members of the LiveMemories project contributing in different ways to the realization of the present work.

[13] http://www.freebase.com
[14] http://dbpedia.org

References

1. F.Ricca, E.Pianta, P.Tonella, and C.Girardi. Improving web site understanding with keyword-based clustering. *Journal of Software Maintenance and Evolution: Research and Practice*, 20(1):1–29, 2008.
2. G.Kobilarov, T.Scott, Y.Raimond, S.Oliver, C.Sizemore, M.Smethurst, C.Bizer, and R.Lee. Media meets semantic web - how the bbc uses dbpedia and linked data to make connections. In *Proceedings of the European Semantic Web Conference (ESWC-2009)*, pages 723–737, Heraklion, Greece, 2009.
3. J.Broekstra, A.Kampman, and F.van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the 1st International Conference on the Semantic Web (ISWC-2002)*, pages 54–68, 2002.
4. J.J.Carroll, C.Bizer P.Hayes, and P.Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th International Conference on World Wide Web (WWW-2005)*, pages 613–622, 2005.
5. Hyuk Jin Ko and Woojun Kang. Enhanced access control with semantic context hierarchy tree for ubiquitous computing. *International Journal of Computer Science and Network Security*, 8(10):114–120, 2008.
6. L.Bentivogli, C.Girardi, and E.Pianta. Creating a gold standard for person cross-document coreference resolution in italian news. In *Proceedings of the Workshop on Resource and Evaluation for Identity Matching, Entity Resolution and Entity Management (LREC-2008)*, pages 19–26, Marrakech, Morocco, 2008.
7. L.Ding, T.Finin, Y.Peng, P.Pinheiro da Silva, and D.L. McGuinness. Tracking rdf graph provenance using rdf molecules. In *Proceedings of the 4th International Semantic Web Conference (ISWC-2005)*, 2005. Poster paper.
8. Hsien-Chou Liao and Chien-Chih Tu. A rdf and owl-based temporal context reasoning model for smart home. *Information Technology Journal*, 6:1130–1138, 2007.
9. M.Benerecetti, P. Bouquet, and C.Ghidini. On the dimensions of context dependence. In P.Bouquet, L.Serafini, and R.H.Thomason, editors, *Perspectives on Contexts*, CSLI Lecture Notes, chapter 1, pages 1–18. Center for the Study of Language and Information/SRI, 2007.
10. M.Speranza. The named entity recognition task at evalita 2009. In *Proceedings of the Workshop Evalita 2009*, Reggio Emilia, Italy, 2009.
11. Matthias Nickles. Social acquisition of ontologies from communication processes. *Appl. Ontol.*, 2(3-4):373–397, 2007.
12. O.Popescu and B.Magnini. Web people search using name entities. In *In Proceedings of the Workshop SemEval-2007*, Prague, CZ, 2009.
13. R.Zanoli, E.Pianta, and C.Giuliano. Named entity recognition through redundancy driven classifiers. In *Proceedings of the Workshop Evalita 2009*, Reggio Emilia, Italy, 2009.
14. T.Stajner, D.Rusu, L.Dali, B.Fortuna, D.Mladenic, and M.Grobelnik. Enrycher : service oriented text enrichment. In *Proceedings of the 11th International multiconference Information Society (IS-2009)*, Ljubljana, Slovenia, 2009.