# Leveraging Entity Linking by Contextualized Background Knowledge

## A case study for news domain in Italian

Andrei Tamilin, Bernardo Magnini, and Luciano Serafini

FBK, Center for Information Technology - IRST
Via Sommarive 18, 38050 Povo di Trento, Italy
{tamilin,magnini,serafini}@fbk.eu

**Abstract.** Entity linking is the process of linking named entity mentions in written texts to entities of background knowledge. On the one hand, information about the context in which entity mentions occurs (e.g., the time, the topic, the geographical location, which the text is relative to) constitutes a critical resource for reducing their ambiguity and for correct entity linking. In fact, without context, mentions are "too little text" to unambiguously refer them to a single entity. On the other hand, in order to profitably exploit contextual information, the background knowledge should be properly organized to support a context-driven access to it. Based on this observation, this paper proposes a context-driven approach to entity linking, by (1) extending standard semantic web knowledge repositories to cope with knowledge contextualized along temporal, thematic, and geographical dimensions, (2) use automatic context detection for text in order to focus the search of entity links in the relevant portions of the knowledge base. The approach has been fully implemented into a working system for processing large-scale news archives in Italian. Assessment of the linking performance has been done using a constructed dataset for linking person entities based on 43K manually annotated news stories of the local Italian newspaper "L'Adige". Experiments demonstrated comparable performance with respect to available entity linking systems for English, and sustained the robustness of the context-driven linking in the presence of ambiguities.

## 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 sources of structured background knowledge in the form of semantic web data, as well as huge amount of trustworthy semi-structured textual data released by various public organizations. Our intuition on how text 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 preexisting background knowledge and (ii) the extracted knowledge extends the background knowledge and as such will contribute to extract new knowledge

from text. In this paper, we focus on phase (i) and show how the use of the information about a *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 mentions, which occur in the text and knowledge they possess on the entities to which mentions refer to (such as people, organization, locations, etc.). The impossibility to create such a link degrades comprehension of a text, while incorrect linking leads to its misunderstanding. Why people can do the linking successfully is by considering *context* in which each entity mention occurs in. For instance, having a newspaper article describing a football match of Italian Professional Football League between Milan and Juventus taken in Turin on 20th of October 2000, reading the *ambiguous* string "Boban", if we already know him, we will correctly link it to the Croatian footballer Zvonimir Boban, who played with Milan team that year, and not, for example, to Boban Marković, the Serbian trumpet player.

In order to turn the depicted human reasoning scenario on exploiting contexts for linking mentions to entities into an automated system, on the one side, there is a need for *machine readable* background knowledge, and on the other side, this knowledge should support *context-driven access* to it. Though, lots of efforts under the semantic web initiative nowadays brought to public vast amounts of background knowledge in from of machine-readable RDF/OWL ontologies, little work has been actually done for representing and managing contextual information bounding the validity and applicability of knowledge.

In order to address these issues and automate entity linking, in this paper we propose a context-driven entity linking approach, which leverages the exploitation of contextual information both by a proper organization of the background knowledge and by enabling the context-driven access to it. To do so, we propose to extend standard semantic web knowledge repository, such as Sesame [5], to cope with knowledge contextualized along temporal, thematic, and geographical dimensions. Formally, we adopted and extended the state of the art *context as a box* framework [12], in which contextual space is defined by a fixed number of partially-ordered dimensions and a concrete context, containing RDF/OWL knowledge base, is defined by a vector of values the corresponding dimensions take. Because of ordered dimensions, contexts in the repository automatically exhibit generalization/specialization ordering allowing to effectively navigate and query contexts. Having the contextualized knowledge repository, the context-driven entity linking consists of the automatic context detection from text in order to focus the search of entity links in the relevant contexts of the repository, possibly shifting search queries to more specific and more general contexts.

For assessment of linking performance we have employed the *accuracy* scoring metric introduced in the Knowledge Base Population campaign at the Text Analysis Conference (TAC KBP) [15]. Due to the focus of the campaign on the processing of English, to enable equivalent evaluation in Italian we have constructed a dataset for linking person entities based on 43K manually annotated news stories of the local Italian newspaper "L'Adige". Evaluation has produced comparable accuracy figures with respect to available entity linking systems for

English participated in the first TAC KBP campaign in 2009, and demonstrated the robustness of the context-driven linking in the presence of ambiguities.

The paper is further structured as follows. In Sect. 2 we define the entity linking task and stress its main difficulties and motivating the use of contexts for overcoming them. The framework for contextual organization of semantic web knowledge bases in further presented in Sect. 3. In Sect. 4 we overview the phase of populating the contextualized repository with actual semantic data. The linking process leveraged by the contextualized knowledge is described in Sect. 5, and further evaluated in Sect. 6. We end up with the overview of related approaches and systems and conclusion.

## 2 Entity Linking

Entity linking is the process of linking mentions of named entities recognized in written texts to corresponding entities contained in a reference background knowledge base, or asserting that the entity is not present in the reference knowledge base. Examples of named entities are person names (e.g., "Zvonimir Boban" referring to Croatian football player), organization names (e.g., "Milan" referring to the football team AC Milan), and location names ("Turin" referring to the Italian city of Turin). Note that an entity (e.g., Zvonimir Boban) can be referred to by different name variants, or surface forms (e.g., "Zvonimir Boban", "Zvonimir", "Boban"). Conversely, some entities can share confusable names, or a surface form can refer to multiple entities (e.g., "Boban" could refer to the football player or to the musician; "Milan" could refer either to the city or to the football team). Both situations create main complications for correct linking.

The entity linking task is defined by a reference background knowledge base on the one side and as the set of *queries* of the following form on the other side:

$$\langle \text{textual document, named entity mention, entity type} \rangle \tag{1}$$

For each query, the linking process either discovers an URI of the referred entity from the reference knowledge base, or it asserts NIL in case there is no corresponding entity in the knowledge base.

For instance, using Wikipedia as the reference knowledge base and given a query consisting of the following document

```
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. After taking a two goal lead from two
headers in two minutes from Ambrosini and Shevchenko, Milan
suffered a soft Trezeguet goal."
```

and the mention "Boban" of type person, the entity linking process should return the corresponding Wikipedia page `http://en.wikipedia.org/wiki/Zvonimir_Boban` about the footballer referred in the text.

In order to correctly complete the above alignment, a linking algorithm should be able to tackle the ambiguity of mention "Boban" by recognizing contextual bounds of the document – that it is about a football match, in 2000 and that the player Zvonimir Boban actually played in it.

Entity linking is related to the task of cross-document co-reference, which is grouping mentions occurring in possibly different texts referring to the very same entity. However, the linking requires alignment to a knowledge base, and not clustering of mentions, hence direct comparison of these two tasks is not supposed to be done.

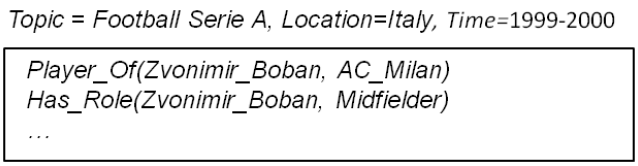
## 3 Contextualized Knowledge Repository

The recognition of the fact that most of the knowledge available on the semantic web in the form of RDF/OWL data is actually tailored to a specific context of use, has recently fostered 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 data provenance and trust of data [10, 6], expressing propositional attitudes [13], dealing with temporally-stamped data [11], and access control [8].

In this section, we present a lightweight *contextualized background knowledge repository* for managing, searching and querying RDF/OWL data qualified with a set of contextual dimensions in the form of temporal, geographic and thematic bounds, those that can be directly derived from input written texts and consequently used for leveraging entity linking algorithms.

### 3.1 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 [12]. 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 Italian Professional Football League Championship, called Serie A, then it can contain the information about participating teams, footballers and the roles they play within teams, etc. Graphically, the corresponding context can be depicted as follows:




Given the original definition of the context as a box framework, we pursue an additional requirement by demanding the values of each of contextual dimensions

to be taken from structured domains, from ontologies, with asserted on them broad-narrow coverage 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.

Practically speaking, the temporal dimension does not require explicit encoding with an ontology and broad-narrow relation for a pair of intervals can be derived on the fly; regarding the geographical dimension, it can be represented by Geonames[1] ontology enumerating geographical places and the geo-political relations between them; and finally for broad-narrow hierarchy of topics one can use DMOZ ontology of topics[2].

The important property of the repository, induced by requirement of hierarchical dimensions, is that we can straightforwardly define the broad-narrow coverage relation between contexts. Given a pair of contexts, $\mathcal{C}_i$ and $\mathcal{C}_j$, defined on the same set of dimensions, we say that $\mathcal{C}_i$ covers $\mathcal{C}_j$ if all dimension values of $\mathcal{C}_i$ cover corresponding dimension values of $\mathcal{C}_j$. Due to that property, contexts of the repository exhibit hierarchical organizations by virtue of values defining them; a newly added context is automatically positioned within the context hierarchy by means of its dimensions values.

### 3.2 Retrieving contextualized knowledge

In answering queries, knowing the scope of the query is of crucial importance. For instance, if one asks for the coach of Milan in the context of Italian Serie A in 2000 or in the context of Italian Serie A nowadays in 2010, he obtains two different answers, i.e. Alberto Zaccheroni and Leonardo Araújo, respectively. From this elementary example one can see how relevant is 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 of the query itself and a (set of) context(s) in which the query should be evaluated.

The choice of the most appropriate context for executing contextualized queries is a very crucial factor. Sending a query to the "wrong" context can produce a wrong or an empty answer. On the other hand, to precisely determine the correct context to 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 can be shifted across contexts. More specifically, as a semantic metric of context closeness, the relation "directly covers" and "is directly covered by" are exploited.

[1] `http://geonames.org`
[2] `http://dmoz.org`

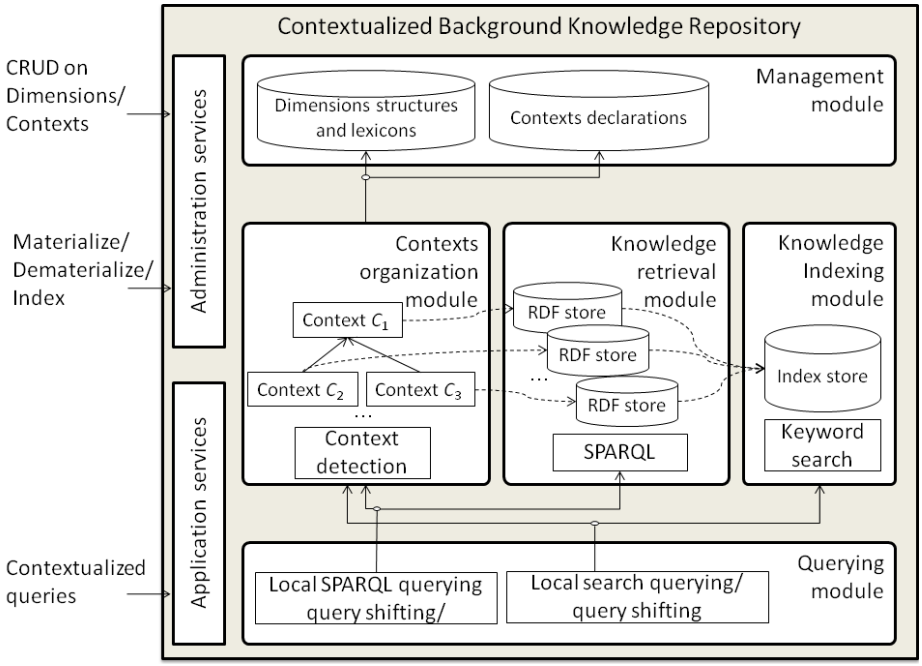



**Fig. 1.** Architecture of the contextualized background knowledge repository

Specification of a (set of) context(s) for a query can be done in two modes. First, by values of contextual dimensions, e.g., $\langle \text{Football}, \text{Italy}, 2009-2010 \rangle$, or by a string of key concepts, e.g., $\langle$"goalkeeper, midfielder", Italy, $2009-2010\rangle$, equivalently matching to the context of the first option. The second option better fits scenarios when the context for a query is detected directly from a text, which is the case study pursued in this paper. In order to implement the context detection operation, lexicons for contextual dimensions have been constructed by associating dimension values with a set of key concepts, relevant for corresponding values. In practice, we only used the lexicon for the topic dimension, selecting as key concepts per value concepts, properties and individuals from terminological part of RDF/OWL knowledge bases contextualized with that value of topic dimension.

### 3.3 Practical implementation

The implementation architecture of the contextualized knowledge repository is graphically depicted in Fig. 1. The repository is composed of five principal modules, whose functionalities are briefly described below.

**Management module** supports the functionalities for (a) defining and managing the dimensions structures and lexicons 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, further used to compute context pairs among which it is possible to shift queries. Context detection operation allows for finding contexts by dimension values or employing matching against dimensions lexicons.

**Knowledge retrieval module** performs the actual loading of knowledge of contexts into the semantic 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-based search queries. In practice, given an RDF/OWL ontology we limited the construction of index from `rdf:ID` and `rdfs:label` elements.

**Querying module** enables to answer contextualized queries using SPARQL and to execute keyword-based search queries over the indexed knowledge. Both types of queries can benefit from query shifting mechanism for diverting queries along the hierarchy of contexts.

Implementation-wise, we grounded our prototype on Sesame RDF/RDFS storage and querying framework [5], which is one of the most popular free open-source tool having good performance and stability. To be able to work with more expressive ontologies, such as ontologies in OWL, we employed the free plugin for the Sesame repository, called SwiftOWLIM[3]. For indexing and text-based searching we used the open-source Apache Lucene Solr tool[4].

## 4 Harvesting Contextualized Background Knowledge

Last years are characterized by a rapid growth of publicly available structured knowledge sets, whose development has been strongly stimulated by recent semantic web initiatives. Among numerous datasets available nowadays, the most notable are Freebase, Wikipedia, DBpedia resources, which provide high quality structured knowledge on well-known or famous entities. Due to the specificity of our case study, which is the analysis of a local Italian news stories, a lot of entities covered in the news stories are missing from those structured sources, though they are contained in various semi-structured forms on the web, such as excel spreadsheets, databases, html pages, etc.

In order to populate the contextualized background knowledge repository we considered both structured and semi-structured sources of knowledge, and developed a number of techniques and procedures for acquiring this knowledge in a context-sensitive way. One of the strong points of the context-dependent knowledge collection is that the acquired knowledge, being inserted in the right context, remains always valid in that context. The repository as a whole might

[3] `http://www.ontotext.com/owlim`
[4] `http://lucene.apache.org/solr/`

not be up-to-date due to some missing knowledge, but the validity of the already inserted knowledge is not violated by the possible changes to the knowledge, for example in time.

The background knowledge acquisition phase in practice resulted in the creation of 700 contexts, containing knowledge about 30K entities (people and organizations). In the following, we present several examples of the performed contextualized knowledge acquisition.

### 4.1 Using semi-structured resources

For acquisition of knowledge from semi-structured textual data released by various official organizations, we have developed a set of semi-automatic procedures for converting to formal RDF/OWL ontologies various data formats, such as database tables, excel sheets, html pages, etc. In particular, we used a number of public classifications of National Statistics Institute (ISTAT)[5], data tables of national and provincial Economical Registries (Camera di Commercio)[6], official web sites of Italian senate and chamber[7], archives of regional and municipal elections of the Autonomous Region of Trentino-Alto Adige[8], archives of sportive events from National Olympic Committee (CONI)[9], etc.). The ontologies has been constructed in Italian language and focused on the following domains of interest: sport (players, teams, regular championships and competitions in football, auto- and moto sport, ice hockey, volleyball, 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[10]).

For generated ontologies, we manually assigned contextual values and imported them into the contextualized knowledge repository. For most of the constructed ontologies, context assignment was rather a straightforward task due to the evident thematic bounds (e.g., different types of sport), temporal bounds (e.g., clear durations of political legislatures), and geographic bounds (e.g., Italy or detailed administrative regions of the Province of Trento) .

### 4.2 Using structured resources

*Importing knowledge from Freebase* Freebase[11] is a massive collaboratively-edited knowledge base of facts about people, organizations, events, etc. Freebase is orga-

[5] `http://www.istat.it/dati/`
[6] `http://www.cameradicommercio.it/`
[7] `http://www.senato.it/`, `http://www.camera.it/`
[8] `http://www.regione.taa.it/elettorale/`
[9] `http://www.coni.it/`
[10] `http://www.trentinocultura.net/territorio/toponomastica/`
[11] `http://freebase.com`

nized into domains (e.g., government, sport disciplines, television, etc.), grouping relevant types (e.g., politicians, political parties, elections, sport championships, players and teams, TV presenters and actors, etc.). Types have properties (e.g., Date of birth for type Person), and can be organized in inheritance hierarchies (e.g., Basketball Player type extends generic type Person) allowing for property inheritance. One of the important points of Freebase, that it supports complex properties, encoding contextually bounded values. An example of such a compound property is Employment history property of Person (inherited by President, Football player, etc.), allowing to state for a person where he worked, and also from/to date limiting temporal validity of the fact.

From Freebase, we mainly harvested non-Italian knowledge regarding famous politicians (mainly presidents and vice-presidents), movie actors, participants of international Olympic events, etc. To identify proper context in our contextualized knowledge repository we used the following strategy: domain defines the topic dimension value, compound properties define temporal interval, and properties connecting facts to Country type define the location dimension. Note that having identified context, generated RDF statements from compound properties of Freebase become simple decontextualized statements.

*Importing knowledge from Italian Wikipedia* The acquisition of knowledge from Italian Wikipedia has been done using special "List of ..." pages grouping people by professions, nationality, temporal periods (e.g., Italian movie actors of XX century). For identification of context, we manually mapped professions to topics in the contextualized repository, Italian nationality to location Italy, and temporal periods to time intervals. Since many pages in Italian Wikipedia do not contain infoboxes, summarizing entity attributes, the acquired knowledge has been limited to the simple statements asserting people's professions.

The acquisition of knowledge from Wikipedia could be automated and simplified by the use of DBpedia [3], however at the moment the content of DBpedia is formed only by the analysis of English Wikipedia pages.

## 5 Context-driven Entity Linking in Italian News Domain

Having on one side news articles, annotated by automated named entity recognizer with textual mentions referring to people, locations and organizations, and the contextualized background knowledge repository populated as presented in Sect. 4, the context-driven entity linking process consists from two principal steps: identification of thematic, temporal and geographic contextual boundaries from texts and identification of a set of relevant contexts in the knowledge repository on the basis of those boundaries; and then exploring the set of identified contexts, possibly shifting to more specific or more general contexts, to search for a match between an input mention and the knowledge contained in the contexts. In the following, we discuss major details of the outlined linking.

### 5.1 Context detection

To define the context of a news article we focused on detection of thematic, temporal and geographic bounds. Thematic bounds capture topics addressed in the article (e.g., generic topics, such as sport, politics, culture, or more specific topics, such as FIFA world cup championship, formula one grand prix, etc.), temporal bounds capture the time boundaries of the article (e.g., the day 2010-06-29, the whole year 2010, the XXI century), and finally geographical bounds capture the geographical region the news in article are bounded to (e.g., national level such as the United States, Italy, or local places, such as Province of Trento, Municipality of Venice, etc.).

In case of arbitrary texts, the task of identifying contextual boundaries is a very hard problem. Fortunately, due to our restriction in the application domain to news articles, some simplifying assumptions can be done due to specificity of news style publishing. Namely, news articles are well-focused (an article as a whole addresses a short number of topics), geographically proximate (prevalent part of news from a local Italian newspaper publishes articles related to local territory it is distributed on), and well-timed (articles contain information relevant to the publication date or near time interval).

Detection of thematic bounds has been casted as selection of a ranked list of keyphrases from a text, describing the most important concepts and providing an approximate but useful characterization of the content of a text. Practically, extraction of keyphrases has been performed using the system KX_FBK demonstrating good performance at the recent keyphrase extraction evaluation campaign at SemEval-2010 [7]. For ranking of extracted keyphrases KX_FBK system uses the standard term frequency - inverse document frequency metric (TF-IDF), combined with a number of additional parameters, such as keyphrase length, position of first occurrence, concept subsumption and others.

Due to the well-timed property of news style, for the identification of temporal bound we adopted the simple strategy of using the publication date of the text, considering for the future more accurate determination of the time from the automatically extracted temporal expressions. For proximity reason, as geographical location bound we by default assigned the value “Italy”, or a municipality name, often explicitly mentioned in the sub header of the news, which is related to a particular territory.

Having identified contextual bounds from a given text, we further employed context determination operation of the contextualized repository to detect ranked list of repository contexts relevant to the given text or state that relevant contexts are missing in the repository. The operation evaluates to matching ranked keyphrases against topic dimension lexicon for finding ranked list of topics from the corresponding dimension, location matches to the proper value in the location dimension, and publication date is used for detecting the time intervals in the repository containing this date.

### 5.2 Linking in a context

Given the identified, non-empty, ranked list $\{\mathcal{C}_i\}_{i=1..n}$ of contexts in the repository, the task of entity linking of a given entity mention $S$ consists in matching string $S$ against the URIs of the knowledge base contained in a context $\mathcal{C}_i$ in the ranked list, possibly shifting to more specific and more general contexts, as we will see later on. Practically, given a context, for matching discovery we employed exact string matching against the indexed elements of the knowledge base in that context. As it has been pointed out in Sect. 3 on the contextualized knowledge repository, for the construction of index of RDF/OWL knowledge base `rdf:ID` and `rdfs:label` elements have been used.

Link discovery procedure has been executed in the loop, evaluating one by one contexts in the ranked list, terminating if in a current context a single matching URI is found and returning it as the result of linking. In case no matching is found, we further apply query shifting operation of the repository. Namely, looping again over contexts in the ranked list, for each context we first detect more specific contexts of it and execute matching in them, terminating in case a single matching URI is discovered, otherwise we detect more general contexts of it and execute again matching in them. In such a way, all ranked contexts are exploited for the linking, as well as more specific and more general contexts are exploited, mitigating possible inaccuracies in identification of ranked contexts.

### 5.3 Practical deployment

The presented context-driven entity linking 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"[12]. 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, etc. Practically, for detection of named entity mentions we used the Typhoon system [16], capable of recognizing in Italian texts named entities referring to people, organizations and locations. 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.

Since the success of entity linking crucially depends on the mention string, e.g., just having a mention "Paolo" does not allow to establish the link to a person Paolo Rossi. That is why before actually applying the linking procedure we first employed co-reference resolution system for clustering mentions referring to the same entity, and afterwards from a cluster, we selected the longest and the most frequent mention, which in practice allows bringing from the text the most representative, complete name, which is the combination of name and surname for people, non abbreviated name for locations and organizations. Using this complete name as input to the linking, we further executed context-driven entity linking wrt. texts the clustered mentions originated from. In practice, for co-reference we used the system developed by Popescu and Magnini [14]. From the

[12] `http://www.ladige.it`

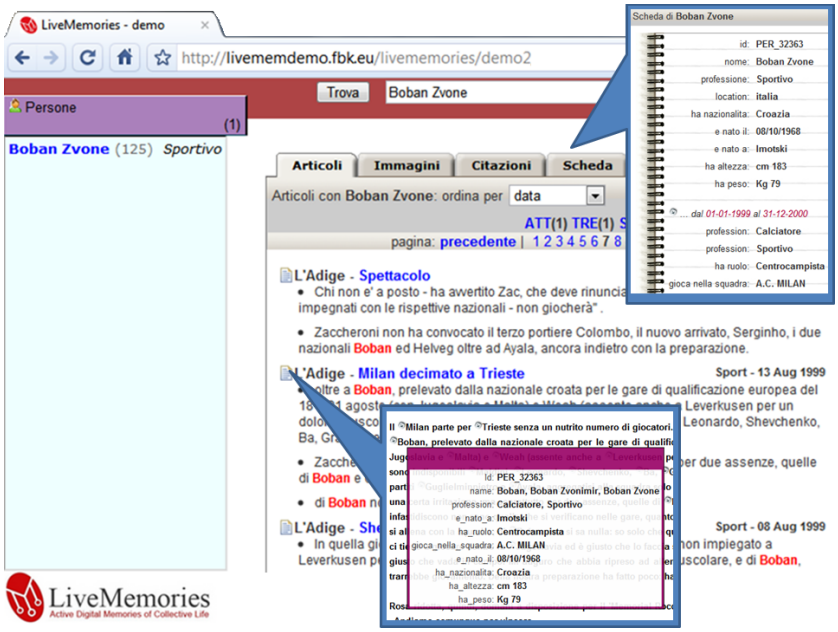



**Fig. 2.** Entity-based news search portal for "L'Adige" archive

extracted named entities of "L'Adige" we further constructed 133,906 clusters corresponding to distinct persons, 34,270 to organizations and 9,836 to locations.

As practical case-study, on top of the processed news archive, we have implemented the news portal, Fig. 2, giving the possibility to search the news archive by entities addressed in the articles, see the contextually relevant background knowledge attached to the linked entities when reading the articles, as well as to see the summarization card displaying all background knowledge on the entity available in the background repository. The functionality of the portal relies on the one side on the context-driven entity linking, and the selection of background knowledge from the contexts in which the linking was performed. This combined operation is called a semantic enrichment [1].

## 6 Experimental Evaluation

In order to evaluate the performance of the context-driven entity linking in the domain of Italian news, we have constructed a data set of manually annotated subset of news articles from the complete "L'Adige" archive, restricted to mentions referring to people. Using the constructed evaluation data, as entity linking scoring metric we employed micro-averaged accuracy, in line with the official evaluation metric of TAC KBP campaign [15], which is defined as following:

$$\text{Accuracy}_{micro} = \frac{\text{Number of queries}}{\text{Total number of queries}} \tag{2}$$

Here the queries have the form (1), as defined in Sect. 2.

In the following, we describe more details on the construction of the evaluation corpus and discuss the results of the experimental evaluation.

### 6.1 Data set

Construction of the linking data set for Italian has been done by reusing existing people cross-document evaluation corpus [9]. The corpus consists of around 43k manually annotated news articles published by the local newspaper "L'Adige" from 1999 to 2006. The annotation has been limited to person entities and to the articles mentioning entities by name. In the majority of cases, annotators used the combination of name and surname, or just a surname, in particularly for famous people, such as in case of Italian singer Zucchero, instead of Zucchero Fornaciari. The corpus contains total number of 209 ambiguous and non-ambiguous names referring to 709 distinct people. For example, ambiguous name "Paolo Rossi" in some articles refers to Italian comic actor Paolo Rossi, in some to football player, in some as a volleyball second coach of the local team Itas Diatec Trentino, in others as local priest, etc. While non-ambiguous name "Zucchero" refers always to the Italian singer `http://www.livememories.org/cultura#zucchero`.

Out of the 709 distinct people contained in the corpus, for 130 people we have manually established the link to corresponding entity in the contextualized knowledge base, others has been verified not to be contained in the knowledge base and hence have been linked to NIL value. For example, Italian comic actor Paolo Rossi has been linked to the URI `http://www.livememories.org/cultura#paolo_rossi`, while a volleyball coach of the local team has been assigned NIL.

In total, as the linking evaluation data set, we have generated 25k queries, among which 17k mentioned entities contained in the contextualized knowledge base and 8k linked to NIL value.

### 6.2 Experiments

As a baseline entity linking method we applied the simple procedure that given a mention searches for a match in the whole contextualized knowledge repository, disregarding the notion of context. In case of multiple matches found the baseline linking has been set to randomly pick a value from matches. This straightforward linking has been evaluated over the developed 25k linked corpus and produced fairly high accuracy figure 0.86. The execution of the context-driven entity linking approach, favoring only the matches in the detected contexts, produced 0.81 accuracy. Thinking about the reasons of having higher figure of the baseline compared to the context-driven linking brought us to the analysis of the contextualized knowledge base and consequent discovery of the fact that actual ambiguity in the knowledge base is fairly low. This consequently puts the baseline not into the random choice of the matches as we have defined it, but rather almost always selecting the single right match. Following this observation we

introduced in the contextualized knowledge repository some additional ambiguities, by inserting randomly in some contexts ambiguous entities, so that given a name in the corpus at least two distinct URIs, i.e., two distinct people with the same name, are present in the repository. Resulting accuracy for the baseline in fact fell down to 0.57 accuracy, while the context-driven linking demonstrated the robustness to the ambiguities producing the 0.78 accuracy.

Despite the differences between evaluating the linking accuracy in TAC KBP evaluation campaign and our evaluation setting, namely, TAC KBP is done for English texts using as the knowledge base English Wikipedia subset, we believe that 0.78 accuracy figure produced by the context-driven entity linking is a positive indicator of good performance given the fact that best performing system for person entity linking in TAC KBP achieved 0.83 accuracy.

## 7 Related Approaches and Systems

*Exploitation of Wikipedia knowledge* for linking named entities is an active research topic, due to the large scale and public availability of the resource. By construction, Wikipedia provides a number of special features allowing for disambiguation process, such as redirect pages, disambiguation pages, as well as categorization of Wikipedia articles. Seminal works of Bunescu and Pasca [2], Cucerzan [17] for leveraging named entity linking by the use of Wikipedia fostered setting up a systematic evaluation campaign for entity linking, namely Knowledge Base Population campaign at the Text Analysis Conference (TAC KBP), first launched in 2009 [15]. A number of systems participated in the first TAC KBP 2009, employing a variety of techniques for linking: casting the problem of entity linking as entity classification using Bayes classifier or as supervised machine learning problem, exploiting different similarity metrics between the document in which the mention occurs in and Wikipedia articles, such as bag of words model for exploitation of information co-occurrence. The performance of the participating systems for linking mentions of type person ranged from the accuracy figures of 0.6 to 0.83 as the best linking system.

*Exploitation of DBpedia knowledge* is another popular approach for entity linking. DBpedia [3] represent the structured excerpt of Wikipedia knowledge. Among numerous works making use DBpedia for linking and disambiguation, we would like to point at two approaches exploiting the notion of context. In [4] the authors present the approach to using DBpedia for interconnecting the data sources within BBC.[13] 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 linking 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

[13] `http://bbc.co.uk`

into corresponding similarity cluster. The work and system Enrycher, described in [18], is another similar example of the completely implemented pipeline for linking textual resources with the knowledge from DBpedia. In a similar vein, it exploits the notion of context in the 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 entity linking is that the background knowledge is organized by construction of the repository into the contexts and the detection of context from a textual resource is based on its geo-temporal and thematic analysis, rather than on co-occurrence of entities.

*Number of industrial-strength systems* for linking entities of unstructured English texts to knowledge base entries has been developed during the last couple of years. Some notable examples are OpenCalais[14] powered by Thomson Reuters, Zemanta[15] blogs linking and enriching with metadata contained in Freebase.

## 8 Conclusion

In this paper, we have presented the entity linking approach leveraged by semantic web knowledge contextualized along thematic, geographic and temporal dimensions. The approach has been fully implemented in the system for linking named entities from the large-scale local Italian news archive "L'Adige". For systematic evaluation of the linking performance, the entity linking data set for Italian has been constructed for linking person entities. Carried on experimental evaluation demonstrated results comparable to existing entity linking systems for English and showed the robustness of the proposed context-driven approach to the scenarios of linking ambiguous named entities, which is when the same name mention refers in different texts, contexts, to different real-world entities.

**Acknowledgements** This work is 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 additionally thank Luisa Bentivogli, Christian Girardi and Roberto Zanoli for valuable discussions on evaluation and assistance with construction of the linking data set.

## References

1. A.Tamilin, B.Magnini, L.Serafini, C.Girardi, M.Joseph, and R.Zanoli. Context-driven semantic enrichment of italian news archive. In *Proceedings of the 7th Extended Semantic Web Conference (ESWC-2010)*, pages 364–378, 2010.
2. R.C. Bunescu and M.Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Conference of Association for Computational Linguistics (EACL-2006)*, 2006.

[14] `http://www.opencalais.com/`
[15] `http://viewer.zemanta.com`

3. C.Bizer, J.Lehmann, G.Kobilarov, S.Auer, C.Becker, R.Cyganiak, and S.Hellmann. Dbpedia a crystallization point for the web of data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, (7):154–165, 2009.
4. 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.
5. 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.
6. 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.
7. S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of SemEval-2010 Workshop at COLING-2010*, Uppsala, Sweden, 2010.
8. 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.
9. L.Bentivogli, C.Girardi, and E.Pianta. Creating a gold standard for person cross-document coreference resolution in italian news. In *Proceedings of LREC-2008*, pages 19–26, Marrakech, Morocco, 2008.
10. 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.
11. 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.
12. 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*, chapter 1, pages 1–18. 2007.
13. Matthias Nickles. Social acquisition of ontologies from communication processes. *Appl. Ontol.*, 2(3-4):373–397, 2007.
14. O.Popescu and B.Magnini. Web people search using name entities. In *In Proceedings of the Workshop SemEval-2007*, Prague, CZ, 2009.
15. P.McNamee and H.Trang Dang. Overview of the tac 2009 knowledge base population track. In *Proceedings of the Text Analysis Conference (TAC-2009)*, Gaithersburg, Maryland, USA, 2009.
16. 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.
17. S.Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL 2007*, page 708716, 2007.
18. 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.