A Logic-based approach to Named-Entity Disambiguation in the Web of Data

Silvia Giannini[1], Simona Colucci(✉)[1],
Francesco M. Donini[2], and Eugenio Di Sciascio[1]

[1] DEI, Politecnico di Bari, Bari, Italy
[2] DISUCOM, Università della Tuscia, Viterbo, Italy

Abstract. Semantic annotation aims at linking parts of rough data (*e.g.*, text, video, or image) to known entities in the Linked Open Data (LOD) space. When several entities could be linked to a given object, a Named-Entity Disambiguation (NED) problem must be solved. While disambiguation has been extensively studied in Natural Language Understanding (NLU), NED is less ambitious—it does not aim to the meaning of a whole phrase, just to correctly link objects to entities—and at the same time more peculiar since the target must be LOD-entities. Inspired by semantic similarity in NLU, this paper illustrates a way to solve disambiguation based on Common Subsumers of pairs of RDF resources related to entities recognized in the text. The inference process proposed for resolving ambiguities leverages on the DBpedia structured semantics. We apply it to a TV-program description enrichment use case, illustrating its potential in correcting errors produced by automatic text annotators (such as errors in assigning entity types and entity URIs), and in extracting a description of the main topics of a text in form of commonalities shared by its entities.

1 Introduction

Web pages represent rich and powerful sources of information collected in form of semi-structured or unstructured text. Semantic technologies offer the possibility to make this knowledge available in a machine processable way, realizing a step forward in the integration and linking of heterogeneous data sources, and enabling different querying mechanisms, involving reasoning and inferences over them. In this context, semantic annotation forms the bridge between textual information and existing ontologies or Semantic Web data sets [1].

Semantic annotation is an Information Extraction (IE) process consisting in applying semantic tags to the information conveyed as free text [3], *i.e.,* it produces a set of Named-Entities (NE) mentioned in the text. It typically requires the recognition of textual spans corresponding to relevant entities, and their classification in a semantic category, possibly assigning a link to real world objects through web identifiers.

In this paper we focus on a specific task within a semantic annotation process: KB-supported entity disambiguation. In brief, it aims at matching a recognized text entity-mention with the corresponding KB entity, resolving polysemy.

Entity disambiguation techniques should also be designed for working in open domains, assuming that there are entities not linkable, *i.e.,* with no match in the reference KB [12]. Disambiguation involves both the identification of the right category of an entity and the choice of the correct entity to link within the KB. In this paper we show how Linked Open Data[3] (LOD) can be successfully used in a disambiguation task, relying on an inference process based on the semantics encoded in the Web of Data. In particular, we present a disambiguation strategy that, starting from a list of DBpedia[4] instances possibly disambiguating an entity, is able to chose the linked web identifier on the basis of a confidence score expressing the specificity level of the information conveyed by that instance when referred to the context defined by other entities mentioned in text. The specificity level is derived with the support of a deductive strategy based on Common Subsumers (CSs) extraction for Resource Description Framework[5] (RDF), for which a definition has been proposed by some of the authors in [9]. CSs are logically computed as RDF descriptions of the features shared by pairs of RDF resources. In this paper we propose a completely novel approach to NE disambiguation, which exploits the informative content embedded in CSs to support the choice—among candidate named entities—of the ones solving the entity linking problem. The adoption of a fully logic-based process—CS extraction—as basis for the proposed strategy makes it logic-based, as well.
The benefit of the proposed approach is illustrated with a possible use-case presented throughout the paper, reporting on the problem of enriching any TV-program description by combining several automatic semantic annotation-tool results.

The paper is organized as follows. Next section gives an overview of the challenges arising in text enrichment through semantic annotations, with particular reference to works related to NE disambiguation. A motivating example is given in Section 3. The formal approach for resolving disambiguation and a proof–of–concept are presented, respectively, in Sections 4 and 5. Conclusions and discussion of future work close the paper.

2 Semantic annotation: an overview

The problem of text enrichment refers to all the tasks involved in extracting useful information from unstructured data. In particular, dealing with semantic annotation of texts spanning over heterogeneous topics—such as TV-programs descriptions—is a challenging goal due to difficulties that arise in training a single extractor able to perform well with texts covering different domains. State-of-the-art NE extractors, including both web services APIs and software frameworks were described by Gangemi *et al.* [15]. They are developed using different algorithms and training data, thus making each of them either targeted for specific NE recognition and classification tasks, or more reliable on particular document

[3] http://linkeddata.org/
[4] http://dbpedia.org/About
[5] http://www.w3.org/RDF/

types (*e.g.,* newspaper articles, scientific papers, etc.) [29]. Therefore, the enrichment of text with heterogeneous topics and formats could benefit from the integration of annotations provided by several extractors [30, 6, 17].

In order to make the problem definition self-contained, we give below an overview of all the processes and challenges to be carried on for producing a valid semantic annotation of a text out of the results provided by different NE extraction tools. The discussion is then focused only on NE disambiguation, for which our contribution is proposed in Sec. 4.

2.1 Named-Entity Recognition (NER)

The first goal to be accomplished by an annotator is to recognize all relevant NE present in a document snippet. As stated in past literature [28, 31], this task is equivalent to recognize all NE that are also keyphrases, discarding irrelevant ones. NER is a challenging problem due to variation in semantically identical but orthographically different entity names (*e.g.,* abbreviations, alternative spellings, or aliases that give rise to different surface forms for the same entity), and presence of entity names with several possible interpretations, that makes the definition of the relevance of an entity context-dependent. Evaluating agreement and disagreement among different extractors results [6] is particularly effective for this process (*e.g.,* an entity missing in one extractor reference KB can be repaired by the others; or, list of equivalent surface forms for the same entity can be obtained evaluating overlapping text spans).

2.2 Named-Entity Classification (NEC)

Given a recognized NE, the classification task requires to assign it to the right semantic category or type. Different supervised or unsupervised methods have been developed in literature for addressing this process [7, 27, 2, 8, 13]. However, an integration strategy of different classifiers outcomes enables the correction of mis-classified entities in terms of category or taxonomy granularity, always selecting the most specific concept representing the class of an entity [6].

2.3 Named-Entity Linking (NEL)

Named-Entity Linking (or Resolution) refers to the association of a NE to the referent entity in a KB. Providing the right resource describing an entity often requires to disambiguate among multiple instances linking to the same entity name or to resolve co-references [23, 11]. The integration of different NE extractors results should improve the enrichment output, repairing missing URIs or correcting wrong links.

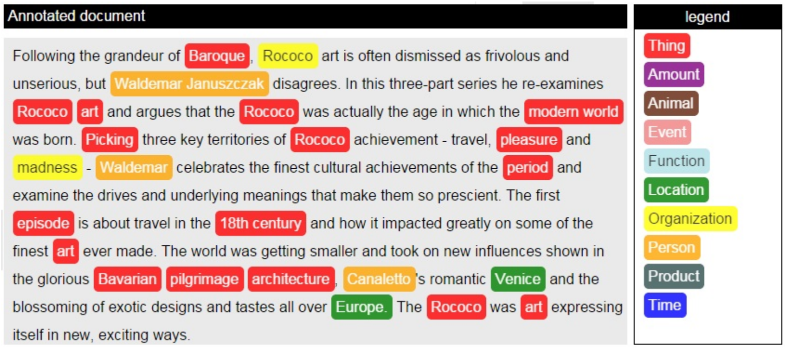
The main contribution of this paper is aimed at the formal definition of a NE disambiguation strategy using the DBpedia KB. Therefore, now we report more in detail on NE disambiguation related work.

Named-Entity Disambiguation (NED) Early approaches to NED were exclusively based on lexical resources, such as WordNet[6] [26]. Linking with encyclopedic knowledge, like Wikipedia[7], then gained more attention due to the larger coverage that these resources can offer in terms of sense inventory for NE and multi-word expressions [25, 21, 4]. Machine learning approaches have been used to identify NE in unstructured text and enrich them with links to appropriate Wikipedia articles [22]. This process, also referred to as wikification, embeds a disambiguation phase relying on a comparison between features (*e.g.,* terms, part–of–speech tags) extracted from the text to be annotated and the Wikipedia pages candidates for disambiguation. Nowadays, the Semantic Web community efforts in structuring and classifying real world entities described in web documents, through fine-grained classification schemes and relational facts in which they are involved, encourage the development of disambiguation techniques based on KB such as DBpedia, Freebase[8] or YAGO[9] [19, 32].
The problem of entity disambiguation has been widely studied for the automatic construction of KBs [24, 12], where entity resolution errors modify the truth value of facts extracted during the KB population process. Hoffart *et al.* ([19]) propose a disambiguation heuristic based on context similarity, where the context is obtained constructing a bag-of-words around the surface name to be linked and each entity the name could be possibly mapped to. The similarity is evaluated considering the word-level overlap between bag-of-words, combined with values expressing the popularity of an entity and the coherence among the context entities. Differently from traditional methods based on bag–of–words for extracting the context, LOD makes implicit relations explicit, thus offering the right structure to exploit pragmatics in disambiguation processes [5], following an idea of resolving interpretation problems with an inference mechanism that is well known in literature [18]. The approach we propose in Section 4 evaluates the most specific information shared by pairs of entities for resolving the disambiguation task, relying on a fully semantic-based exploration of the RDF descriptions of the involved resources.

3 A motivating example

Consider the text in Fig.1 representing the description of the first episode of the BBC TV-series *Rococo: Travel, Pleasure, Madness*[10]. As a baseline motivating our work, we report in Tab.2 the annotation results provided by the tool NERD[11]. Each NE is anchored to the corresponding text phrase, labeled with a type of the NERD ontology, and linked to a web page describing the entity

[6] http://wordnet.princeton.edu/
[7] http://www.wikipedia.org/
[8] http://www.freebase.com/
[9] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[10] http://www.bbc.co.uk/programmes/b03sg830#programme
[11] http://nerd.eurecom.fr/analysis#



Annotated document

Following the grandeur of Baroque, Rococo art is often dismissed as frivolous and unserious, but Waldemar Januszczak disagrees. In this three-part series he re-examines Rococo art and argues that the Rococo was actually the age in which the modern world was born. Picking three key territories of Rococo achievement - travel, pleasure and madness - Waldemar celebrates the finest cultural achievements of the period and examine the drives and underlying meanings that make them so prescient. The first episode is about travel in the 18th century and how it impacted greatly on some of the finest art ever made. The world was getting smaller and took on new influences shown in the glorious Bavarian pilgrimage architecture, Canaletto's romantic Venice and the blossoming of exotic designs and tastes all over Europe. The Rococo was art expressing itself in new, exciting ways.

Fig. 1. Annotation output produced by the tool NERD for the first-episode synopsis of the BBC TV-series *Rococo: Travel, Pleasure, Madness*. All recognized NE are highlighted in the annotated document, with different colors depending on the category assigned in NERD. Category colors are explained in the legend on the right. For the list of assigned URIs see Tab. 2.

(all prefixes introduced throughout the paper are listed in Tab. 1). NERD is

dbpedia:	< http : //dbpedia.org/resource/ > .
dbpedia-owl:	< http : //dbpedia.org/ontology/ > .
dc-terms:	< http : //purl.org/dc/terms/ > .
foaf:	< http : //xmlns.com/foaf/0.1/ > .
all-art:	< http : //www.all-art.org/history252_contents_ > .
nerd:	< http : //nerd.eurecom.fr/ontology# > .
wikipedia:	< http : //en.wikipedia.org/wiki/ > .
skos:	< http : //www.w3.org/2004/02/skos/core# >

Table 1. Prefixes used in the examples of this paper

a web framework [30] that provides a set of axioms for aligning taxonomies of several other NE extractors with the one present in NERD, and a machine-learning based strategy for combining different extractor outputs with conflicts resolution. We highlight some kind of errors still contained in the outcome of the aggregation process for the NEL task. In particular, a distinction between mapping errors (the URIs in italic font in Tab. 2) and disambiguation errors (the ones in bold font) can be made. The former are represented by links that are not DBpedia resources (the AlchemyAPI[12], Extractiv[13], OpenCalais[14], TextRa-

[12] http://www.alchemyapi.com/
[13] http://www.programmableweb.com/api/extractiv
[14] http://www.opencalais.com/

zor[15], Wikimeta[16], Yahoo! Content Analysis[17], and Zemanta[18] extractors integrated in NERD return links to Wikipedia pages) and are easily repairable (the DBpedia resource corresponding to a Wikipedia page can be identified through the property `foaf : primaryTopic`). Links to DBpedia redirection pages, consisting of redirection hyperlinks from alternative names to the actual resource describing the entity, can also be treated as mapping errors. On the other hand, disambiguation errors are constituted by links to wrong resources (Rows 11, 13, 17), or by references to DBpedia disambiguation pages (Rows 7, 14), that are not useful URIs for the linking task. This kind of error seems to be influenced by the type assigned to the corresponding NE (see the `nerd : Organization` type in Row 11 and 17). Moreover, missing links are also present, denoting a bad co-reference resolution process or the recognition of not relevant NE.

In the next section, we describe an approach for resolving ambiguity in NE linking. The approach is based on an inference service specifically defined for RDF.

4 NE disambiguation: a deductive approach

4.1 Problem definition

Consider a text as a sequence of tokens $\mathcal{T} = \langle t_1, \ldots, t_n \rangle$, where n is the number of words in $\mathcal{T}$. A knowledge-based semantic-annotation tool extracts a set of entities $\mathcal{E} = \{e_1, \ldots, e_m\}$ from $\mathcal{T}$, where each entity e_i, $i \in \{1, \ldots, m\}$, is identified by a tuple $e_i = (\mathit{label}_i, \mathit{offset}_i, \mathit{type}_i, \mathit{URI}_i)$[22], where label_i is a sequence $\langle t_j, \ldots, t_k \rangle$ of tokens in $\mathcal{T}$ s.t. $1 \leq j < k \leq n$ and $\mathit{label}_i \cap \mathit{label}_{i+1} = \emptyset$, $\forall i \in \{1, \ldots, m-1\}$, and offset_i is the position of the first character of label_i in the text. With URI_i and type_i we indicate a set of resources, and corresponding types, identified in the KB as possible annotation entities for label_i. Entity e_i is defined *unambiguous* if it results $|\mathit{URI}_i| \in \{0, 1\}$ and, consequently, $|\mathit{type}_i| \in \{0, 1\}$; otherwise, it is called *ambiguous*. Here, we illustrate a completely novel method for resolving URI conflicts, and possibly the related type conflicts, in presence of ambiguous entities, leveraging on the DBpedia KB. The disambiguation process is grounded on the fully logic–based extraction of commonalities between pairs of RDF resources, through the computation of Common Subsumers (CS). Although adopting an already presented algorithm for such a computation [9], the proposed approach to disambiguation is completely novel.

[15] https://www.textrazor.com/

[16] http://www.wikimeta.com

[17] https://developer.yahoo.com/search/content/V2/contentAnalysis.html

[18] http://developer.zemanta.com/

[19] http://spotlight.dbpedia.org/

[20] https://ner.vse.cz/thd/

[21] http://www.old.ontotext.com/lupedia

[22] Please note that semantic annotation systems often return also a confidence value for the extracted NE denoting the probability for that entity to be correctly annotated by the given type and URI.

Table 2. List of entity label, offset, type, and URI for each NE returned by the NERD framework for the text snippet shown in Fig. 1 (the extractor producing each NE is also reported in the last column).

ID	Label name	Offset	Type	URI	Extractor
1	18th century	533	`nerd : Thing`	`dbpedia : 18th_century`	DBpedia Spotlight[19]
2	Architecture	710	`nerd : Thing`	`dbpedia : Architecture`	TextRazor
3	Art	176	`nerd : Thing`	*wikipedia : Art*	TextRazor
4	Art	598	`nerd : Thing`	*wikipedia : Art*	TextRazor
5	Art	832	`nerd : Thing`	*wikipedia : Art*	TextRazor
6	Baroque	26	`nerd : Thing`	`dbpedia : Baroque`	Wikimeta
7	Bavarian	690	`nerd : Thing`	**dbpedia : Bavarian**	THD[20]
8	Canaletto	724	`nerd : Person`	`dbpedia : Canaletto`	Wikimeta
9	Episode	502	`nerd : Thing`	`dbpedia : Episode`	TextRazor
10	Europe	809	`nerd : Location`	`dbpedia : Europe`	DBpedia Spotlight
11	Madness	339	`nerd : Organization`	**dbpedia : Madness_(band)**	DBpedia Spotlight
12	Modern world	241	`nerd : Thing`	`dbpedia : Modern_history`	DBpedia Spotlight
13	Period	409	`nerd : Thing`	**dbpedia : Period_piece**	DBpedia Spotlight
14	Picking	264	`nerd : Thing`	**dbpedia : Picking**	THD
15	Pilgrimage	699	`nerd : Thing`	`dbpedia : Pilgrimage`	Wikimeta
16	Pleasure	326	`nerd : Thing`	`dbpedia : Pleasure`	TextRazor
17	Rococo	35	`nerd : Organization`	**dbpedia : Rococo_(band)**	Lupedia[21]
18	Rococo	169	`nerd : Thing`	`dbpedia : Rococo`	Wikimeta
19	Rococo	200	`nerd : Thing`	`dbpedia : Rococo`	Wikimeta
20	Rococo	297	`nerd : Thing`	`dbpedia : Rococo`	Wikimeta
21	Rococo	821	`nerd : Thing`	`dbpedia : Rococo`	Wikimeta
22	Venice	745	`nerd : Location`	`dbpedia : Venice`	Wikimeta
23	Waldemar	349	`nerd : Person`		Wikimeta
24	Waldemar Januszczak	97	`nerd : Person`	`dbpedia : Waldemar_Januszczak`	Wikimeta

We argue that the informative content determining the context of an entity e_i is conveyed by the entity itself and a set of neighbor entities, expressible though a sliding window $[e_{i-j}, e_{i+k}]$. Given an ambiguous entity e_i, indices $j < i$ and $k > i$ can be respectively decremented and incremented until meaningful commonalities between the RDF description of a possible URI_i and the URI of another entity in the window are not found by the CS algorithm, resolving ambiguity. Moreover, the method we propose extracts a description of the context shared by neighbor entities in terms of their CS, thus paving the way also to other types of IE tasks (*e.g,* topic extraction [14]). We also remark that in this paper we do not deal with the identification of candidate entities, but we rely on results provided by existing semantic-annotation tools.

4.2 Algorithm description

The adoption of a deductive approach allows for investigating on the informative content hidden in the input resources in order to find out the features shared among them and to ground the disambiguation on the evaluation of such commonalities. In other words, CSs are analyzed to discover in what two RDF resources are similar and consequently to rank pairs of resources according to the significance of the informative content they share. Here we just sketch the distinguishing features of the adopted logic-based approach for computing a CS of two RDF resources. The reader interested in more details about the computation algorithm may refer to Colucci *et al.* [9].

Given an ambiguous entity e_i, for each $uri_i^k \in URI_i$, $k \in \{1, \ldots, |URI_i|\}$, the Common Subsumer (CS) extraction algorithm is run, taking as input the pair $uri_i^k, \overline{uri}_j$, where $\overline{uri}_j$ is the URI of a unambiguous selected entity in the neighborhood of e_i.

The strategy heavily relies on the representation of the resources input to the CS computation. In fact, the meaning of an RDF resource r changes depending which triples r is involved in, since different sets of triples entail (in general) different new triples. The approach by Colucci *et al.* proposes to consider a set of triples T_r for resource r, defining the pair $\langle r, T_r \rangle$ as a rooted-graph (r-graph). The criterion for computing T_r is flexible, allowing a user for setting several selection parameters, such as the exploration level d of the RDF graph rooted in r, the datasets to be used as information sources for triples selection and the properties to be included in the chosen triples.

In brief, given an exploration depth d, the algorithm performs a process similar to a depth-first exploration of the two r-graphs describing resources uri_i^k and $\overline{uri}_j$. Namely, the CS of a pair $(\langle r, T_r \rangle, \langle s, T_s \rangle)$ is $\langle r, T_r \rangle$ itself if $\langle r, T_r \rangle = \langle s, T_s \rangle$; otherwise, it is represented by a pair $\langle \_ : x, T \rangle$, with $\_ : x$ a blank node and T a set of triples entailed (according to Simple Entailment [16]) during the parallel exploration of the two r-graphs. Moreover, the traditional process of depth-first exploration is changed to also explore RDF triples describing predicates encountered in previously investigated triples. Accordingly, the algorithm recursively computes the CS by comparing all resources (both predicates and objects) at the same level in the parallel depth-first search of the two r-graphs. At the end,

it results that each triple in T is entailed by a pair of triples encountered in the joint exploration of the two input r-graphs. By example, given a CS $\langle \_ : x, T \rangle$ of a pair of RDF resources $(\langle r, T_r \rangle, \langle s, T_s \rangle)$ and an investigation distance $d = 1$, triples in T are of the form $t = \_ : x\ y\ z.$
If such a triple belongs to T, then the two triples $t_1 = r\ p_1\ q_1.$ (with $t_1 \in T_r$), and $t_2 = s\ p_2\ q_2.$ (with $t_2 \in T_s$), both exist and both entail t. By looking at t, we notice that $\_ : x$ is the root of the CS and y and z may be either URIs or blank nodes. If y (respectively z) is a URI (or a Literal value for z), $y = p_1 = q_1$ (respectively $z = p_2 = q_2$[23]).

4.3 Ranking Common Subsumers

The CS still comes in an r-graph format which includes an RDF description: the result set T. Therefore, we need here to define metrics for identifying the best candidate URI resolving the disambiguation task. Such metrics, which make part of the main contribution of this paper, quantify the amount of information embedded in CS description, and therefore shared by the two initial RDF resources.

In most ambiguous cases, disambiguation involves choosing a *sequence* of entities—two or more, not a single one—for a sequence of labels. In some cases, entities in the "right" disambiguation share a common context—some specific description—that their CS can make explicit. In contrast, wrong choices assign entities which do not share a common context, and this fact can be detected because their CS is some very generic description. We propose to resolve these cases by extracting and comparing the CSs of each candidate sequence of entities—in this paper, we limit to pairs. To this end, we provide a function converting the informative content conveyed by the CS in a score estimating how much specific is the common context of the two resources. Such a score is useful for establishing the best candidate URI resolving the disambiguation task.

Intuitively, the triples $t = \_ : x\ y\ z.$ in T which make a CS more significant are the ones in which both y and z are identifiable resources—*i.e.,* they are not blank nodes. As a consequence, the ranking function is designed in order to assign a weight equal to 1 to such triples. Triples t in which only the object is defined (*i.e.*, z is a non-anonymous resource), are ranked with a lower weight, 0.8. Finally, if a triple t has only a definite predicate (*i.e.,* y is not a blank node and the object z is an anonymous resource), the ranking function gives a weight 0.2 to t. Therefore, the triple ranking function is represented by the following relation:

$$rf(t) = rf_p(t) + rf_o(t), \tag{1}$$

where $rf_p(t)$ and $rf_o(t)$ are, respectively, the functions ranking the predicate and the object of triple t, defined as

$$rf_p(t) = \begin{cases} 0 & \text{if } p \text{ is a blank node} \\ 0.2 & \text{otherwise} \end{cases} ; \tag{2}$$

[23] This relation is also valid when q_1 and q_2 are Literal nodes.

$$rf_o(t) = \begin{cases} 0 & \text{if } o \text{ is a blank node} \\ 0.8 & \text{otherwise} \end{cases} . \tag{3}$$

Then, the metrics expressing how much r and s have in common on the basis of the significance of their CS $\langle \_ : x, T \rangle$ is given by

$$rank(r, s) = \frac{\sum_{t \in T} rf(t)}{|T|} \tag{4}$$

It is worth noting that the minimum value assumed by $rank(r, s)$ is 0.2 and it is assigned to pair of resources sharing only some predicates of the data model. The maximum value, that is 1, is instead obtained when $r = s$ holds.

5 Proof of concept

For illustrating the disambiguation methodology, we consider the first wrong entity of the text in Fig. 1, which is (Rococo, 35, `nerd : Organization`, `dbpedia : Rococo_(band)`). In order to obtain a list of candidate URIs for the repairing process, the DBpedia endpoint[24] is queried for the resource `dbpedia : Rococo_(disambiguation)`[25]. Table 3 shows the CS triples and the value of the CS ranking function for all possible pairs of resources constituted by `dbpedia : Baroque` as first item and one of the disambiguating URIs listed in `dbpedia : Rococo_(disambiguation)` as second item (the exploration depth d is set to 1). Resource `dbpedia : Rococo`, obtaining the highest specificity score, is correctly selected by the disambiguation method. Please note that the second result is still pertaining to the Decorative Art macro-category, but less specific. A side advantage of the proposed disambiguation method is the correction of classification errors for entities with a disambiguation error in the NEL phase (the entity considered in the example was classified as Organization by NERD).

We evaluated also the performance of the approach in the unlucky case that also entity (Baroque, 26, `nerd : Thing`, `dbpedia : Baroque`), previously considered unambiguous, is affected by a disambiguation error in the semantic annotation process[26]. Values obtained by the ranking function for the 440 pairs of resources disambiguating entities Baroque and Rococo are reported at http://193.204.59.20/ned/baroque_rococo.html. The highest value of specificity (0.9) is obtained by pair ⟨`dbpedia : Baroque_Architecture`, `dbpedia : Rococo`⟩, while the first result belonging to the Music category obtains a 0.74 value for ⟨`dbpedia : Baroque_orchestra`, `dbpedia : Variations_on_a_Rococo_Theme`⟩. In this case, other neighboring entities can be considered for further defining the context and disambiguate clearly the meaning of the two resources (*e.g.,* through a sequence of

[24] http://dbpedia.org/sparql

[25] Alternatively, a query searching for all resources having object literal values of properties `rdfs : label` or `foaf : name` similar to the label of the ambiguous entity can be set up.

[26] The interested reader could verify through the Lupedia web service that this extractor annotates Baroque and Rococo respectively with the wrong entities `dbpedia : Baroque_(band)` and `dbpedia : Rococo_(band)`

rank	<dbpedia : Baroque, dbpedia : Rococo>

0.84	_ : cs0 dc − terms : subject dbpedia : Category : 18th_century_in_art, dbpedia : Category : Decorative_arts, dbpedia : Category : Early_Modern_period, _ : z1;
	_ : y1 dbpedia : Category : 18th_century_in_art, dbpedia : Category : Decorative_arts, dbpedia : Category : Early_Modern_period;
	dbpedia − owl : wikiPageExternalLink all − art − history : Baroque_Rococo.html, _ : z2;
	dbpedia − owl : wikiPageWikiLink dbpedia : Category : 18th_century_in_art, dbpedia : Category : Decorative_arts, dbpedia : Category : Early_Modern_period, dbpedia : History_painting, dbpedia : Victoria_and_Albert_Museum, dbpedia : Art_history, dbpedia : Protestant_Reformation, dbpedia : William_Kent, dbpedia : Counter − Reformation, dbpedia : Jean − Philippe_Rameau, dbpedia : Neoclassicism, dbpedia : History_of_wood_carving, dbpedia : Palladian_architecture, dbpedia : Aleijadinho, _ : z3;
	foaf : isPrimaryTopicOf _ : z4.

rank	**<dbpedia : Baroque, dbpedia : Rococo_Revival>**

0.60	_ : cs0 dc − terms : subject dbpedia : Category : Decorative_arts, _ : z1;
	_ : y1 dbpedia : Category : Decorative_arts;
	dbpedia − owl : wikiPageWikiLink dbpedia : Category : Decorative_arts, _ : z2;
	foaf : isPrimaryTopicOf _ : z3.

rank	**<dbpedia : Baroque, dbpedia : Variations_on_a_Rococo_Theme>**

0.47	_ : cs0 dc − terms : subject _ : z1;
	dbpedia − owl : wikiPageExternalLink _ : z2;
	dbpedia − owl : wikiPageWikiLink dbpedia : Concerto, dbpedia : Rococo, _ : z3;
	foaf : isPrimaryTopicOf _ : z4.

rank	**<dbpedia : Baroque, dbpedia : Rocky_Rococo>**

0.20	_ : cs0 dc − terms : subject _ : z1;
	dbpedia − owl : wikiPageExternalLink _ : z2;
	dbpedia − owl : wikiPageWikiLink _ : z3;
	foaf : isPrimaryTopicOf _ : z4.

rank	**<dbpedia : Baroque, dbpedia : Rococo_(band)>**

0.20	_ : cs0 dc − terms : subject _ : z1;
	dbpedia − owl : wikiPageExternalLink _ : z2;
	dbpedia − owl : wikiPageWikiLink _ : z3;
	foaf : isPrimaryTopicOf _ : z4.

rank	**<dbpedia : Baroque, dbpedia : Rococo_(club)>**

0.20	_ : cs0 dc − terms : subject _ : z1;
	dbpedia − owl : wikiPageExternalLink _ : z2;
	dbpedia − owl : wikiPageWikiLink _ : z3;
	foaf : isPrimaryTopicOf _ : z4.

Table 3. CSs enumeration for correcting the annotation of entity (Rococo, 35, `nerd : Organization`, `dbpedia : Rococo_(band)`) ordered by values of the ranking function (each evaluated resources pair is reported in bold).

skos : broader predicates, it is possible to reach the two categories dbpedia : Category : Aesthetics and dbpedia : Category : Visual_arts that are also the subject of the recognized linked entity dbpedia : Art). For what concerns the refinement of the entity linking process for dbpedia : Picking, dbpedia : Madness_(band) and dbpedia : Period_piece resources, the computed rank value is 0.2 for the CSs of all possible combination of disambiguation URIs with entities dbpedia : Rococo and dbpedia : Pleasure of the same sentence. It means that the reference KB does not contain any instance representing those entities. Moreover, the reader may agree that the considered ambiguous entities are not relevant, apart from the one labeled with Madness, whose relevance is gained by its occurrence in the title.

Finally, the last unresolved entity is (Bavarian, 690, nerd : Thing, dbpedia : Bavarian). Here, the proposed method returns the highest value, that is 0.57, for the pair ⟨dbpedia : Bavarian, _Iran, dbpedia : Europe⟩, due to the peculiarity of being both of type dbpedia-owl : PopulatedPlace. Future work will be spent on managing these cases, introducing a threshold on the CS ranking values or adopting other NE techniques to support the same task.

6 Conclusions and Future Work

This paper addresses the most challenging aspects in semantic annotation processes and presents the formal definition of a method for resolving NE disambiguation. We show that realizing disambiguation by evaluating commonalities of entities contained in a text powerfully leads to the identification of extensional-based alternative contexts, and the evaluation of the degrees of specificity of the identified contexts drives the disambiguation process. This paper is far from being exhaustive. Extensive studies on the effectiveness of the approach compared with other techniques already proposed in the literature are under analysis. In particular, performance on linking accuracy, in terms of known and unknown KB-entity correctly linked, will be evaluated. Performances after the inclusion of other LOD datasets have also to be studied. Moreover, meaningful CSs descriptions could be used for retrieving other resources conveying the same informative content of some NE in the text [10], providing a support for serendipitous encounters [22, 20].

Acknowledgments

We acknowledge support of project "A Knowledge based Holistic Integrated Research Approach" (KHIRA - PON 02_00563_3446857).

References

1. New dimensions in semantic knowledge management
2. Alfonseca, E., Manandhar, S.: An unsupervised method for general named entity recognition and automated concept discovery. In: Proc. of the 1st Int. Conf. on General WordNet, Mysore, India. pp. 34–43 (2002)
3. Bellot, P., Bonnefoy, L., Bouvier, V., Duvert, F., Kim, Y.M.: Large Scale Text Mining Approaches for Information Retrieval and Extraction. In: Innovations in Intelligent Machines-4, pp. 3–45. Springer (2014)
4. Bunescu, R.C., Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation. In: Proc. of the 11th Conf. of the European Chapter of the Association for Computational Linguistics (EACL-06). vol. 6, pp. 9–16 (2006)
5. Cambria, E., White, B.: Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine 9(2), 48–57 (2014)
6. Chen, L., Ortona, S., Orsi, G., Benedikt, M.: Aggregating semantic annotators. Proc. of the VLDB Endowment 6(13), 1486–1497 (2013)
7. Chieu, H.L., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: Proc. of the 19th Int. Conf. on Computational linguistics-Volume 1. pp. 1–7. ACL (2002)
8. Cimiano, P., Völker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP) (2005)
9. Colucci, S., Donini, F.M., Di Sciascio, E.: Common Subsumbers in RDF. In: Proc. of AI* IA 2013: Advances in Artificial Intelligence. pp. 348–359. Springer (2013)
10. Colucci, S., Giannini, S., Donini, F.M., Di Sciascio, E.: A deductive approach to the identification and description of clusters in Linked Open Data. In: Proc. of the 21th European Conf. on Artificial Intelligence (ECAI '14). IOS Press (2014)
11. Cucerzan, S.: TAC entity linking by performing full-document entity extraction and disambiguation. In: Proc. of the Text Analysis Conference. vol. 2011 (2011)
12. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proc. of the 23rd Int. Conf. on Computational Linguistics. pp. 277–285. ACL, Beijing, China (August 2010)
13. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence 165(1), 91–134 (2005)
14. Fetahu, B., Dietze, S., Pereira Nunes, B., Antonio Casanova, M., Taibi, D., Nejdl, W.: What's all the data about?: creating structured profiles of linked data on the web. In: Proc. of the companion publication of the 23rd Int. Conf. on World wide web companion. pp. 261–262. International World Wide Web Conferences Steering Committee (2014)
15. Gangemi, A.: A comparison of knowledge extraction tools for the Semantic Web. In: The Semantic Web: Semantics and Big Data, pp. 351–366. Springer (2013)
16. Hayes, P.: RDF Semantics, W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-mt-20040210/ (2004)
17. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using Linked Data. In: The Semantic Web–ISWC 2013, pp. 98–113. Springer (2013)
18. Hobbs, J.R., Stickel, M., Martin, P., Edwards, D.: Interpretation as abduction. In: Proc. of the 26th annual meeting on Association for Computational Linguistics. pp. 95–103. ACL (1988)

19. Hoffart, J., Yosef, M.A., Bordino, Ilaria Fürstenau and, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. pp. 782–792. ACL, Edinburgh, UK (July 2011)
20. Maccatrozzo, V.: Burst the filter bubble: using semantic web to enable serendipity. In: The Semantic Web–ISWC 2012, pp. 391–398. Springer (2012)
21. Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: Proc. of the 16th ACM Conf. on Information and Knowledge Management. pp. 233–242. ACM (2007)
22. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management. pp. 509–518. ACM (2008)
23. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics 2, 231–244 (2014)
24. Nakashole, N., Theobald, M., Weikum, G.: Scalable knowledge harvesting with high precision and high recall. In: Proc. of the fourth ACM Int. Conf. on Web search and data mining. pp. 227–236. ACM, Hong Kong (February 2011)
25. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
26. Navigli, R., Velardi, P.: Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(7), 1075–1086 (2005)
27. Niu, C., Li, W., Ding, J., Srihari, R.K.: A bootstrapping approach to named entity classification using successive learners. In: Proc. of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. pp. 335–342. ACL (2003)
28. Prokofyev, R., Demartini, G., Cudré-Mauroux, P.: Effective named entity recognition for idiosyncratic web collections. In: Proc. of the 23rd Int. Conf. on World wide web. pp. 397–408. International World Wide Web Conferences Steering Committee (2014)
29. Rizzo, G., Erp, M.V., Troncy, R.: Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In: Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (may 2014)
30. Van Erp, M., Rizzo, G., Troncy, R.: Learning with the Web: Spotting Named Entities on the Intersection of NERD and Machine Learning. In: # MSM. pp. 27–30. Citeseer (2013)
31. Zhang, L., Pan, Y., Zhang, T.: Focused named entity recognition using machine learning. In: Proc. of the 27th annual Int. ACM SIGIR Conf. on Research and development in information retrieval. pp. 281–288. ACM (2004)
32. Zheng, Z., Si, X., Li, F., Chang, E.Y., Zhu, X.: Entity disambiguation with freebase. In: Proc. of the The 2012 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology-Volume 01. pp. 82–89. IEEE Computer Society (2012)