HAL
archives-ouvertes.fr

REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets

Carmen Brando, Francesca Frontini, Jean-Gabriel Ganascia

▶ To cite this version:

Carmen Brando, Francesca Frontini, Jean-Gabriel Ganascia. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. Complex Systems Informatics and Modeling Quarterly, 2016, pp.60 - 80. <https://csimq-journals.rtu.lv/>. <10.7250/csimq.2016-7.04>. <hal-01396037>

HAL Id: hal-01396037
https://hal.sorbonne-universite.fr/hal-01396037

Submitted on 13 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Complex Systems Informatics and Modeling Quarterly
CSIMQ, Issue 7, June/July 2016, Pages 60–80
Published online by RTU Press, https://csimq-journals.rtu.lv
http://dx.doi.org/10.7250/csimq.2016-7.04
ISSN: 2255-9922 online

REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets

Carmen Brando[1], Francesca Frontini[2] and Jean-Gabriel Ganascia[3]

[1]Centre de Recherches Historiques, School for Advanced Studies in the Social Sciences (EHESS), UMR 8558, 190-198 avenue de France, 75013 Paris, France
[2]Istituto di Linguistica Computazionale "Antonio Zampolli", Consiglio Nazionale delle Ricerche, Area della Ricerca di Pisa, Via Giuseppe Moruzzi No 1, 56124 Pisa, Italy
[3]Labex Observatoire de la vie littéraire (OBVIL). Laboratoire d'Informatique de Paris 6 (LIP6), Pierre and Marie Curie University, UMR 7606, 4 place Jussieu, 75005, Paris, France

Carmen.Brando@ehess.fr, Francesca.Frontini@ilc.cnr.it, Jean-Gabriel.Ganascia@lip6.fr

Abstract. This paper proposes a graph-based Named Entity Linking (NEL) algorithm named REDEN for the disambiguation of authors' names in French literary criticism texts and scientific essays from the 19th and early 20th centuries. The algorithm is described and evaluated according to the two phases of NEL as reported in current state of the art, namely, candidate retrieval and candidate selection. REDEN leverages knowledge from different Linked Data sources in order to select candidates for each author mention, subsequently crawls data from other Linked Data sets using equivalence links (e.g., owl:sameAs), and, finally, fuses graphs of homologous individuals into a non-redundant graph well-suited for graph centrality calculation; the resulting graph is used for choosing the best referent. The REDEN algorithm is distributed in open-source and follows current standards in digital editions (TEI) and semantic Web (RDF). Its integration into an editorial workflow of digital editions in Digital humanities and cultural heritage projects is entirely plausible. Experiments are conducted along with the corresponding error analysis in order to test our approach and to help us to study the weaknesses and strengths of our algorithm, thereby to further improvements of REDEN.

Keywords: Named Entity Linking, graph centrality, linked data, data fusion, digital humanities.

1 Introduction

To discover new information and to compare it to other sources of information are two important 'scholarly primitives', basic activities common to research across humanities disciplines [1], and especially to those which involve the study of textual sources. Within the Digital Humanities (DH), several instruments have been developed in order to facilitate the work of scholars in these tasks. In particular, the XML-based Text Encoding Initiative (TEI) standard[1] [2] for digital editions allows for the explicit encoding of information in texts, so that they become machine readable and searchable. Furthermore, XML-TEI enables the semantic enrichment of texts, namely, the annotation of portions of texts with tags that connect them with other sources of information, that are not present in the original text. Typically, a word representing a concept, or an entity, or a fact mentioned in a text can be connected to an external link containing further information about them.

[1] http://www.tei-c.org/index.xml

Semantic annotation is not only about an enhanced reading experience. In fact, if the target of the link contains structured, machine readable information, then semantically enriched texts can be processed and analysed in a non-linear and automatic way, discovering connections between different (parts of) texts, aggregating data, comparing and visualising it. Clearly, the production of quality digital editions is not an easy task, and requires manual annotation and validation. Nevertheless, Natural Language Processing (NLP) tools are often used to speed up the process to a great extent.

This paper is set in the broader context of NLP tools for the semantic annotation of Named Entity (NE) mentions, and in particular of mentions of places, persons and organisations in digital editions in the literary domain. More specifically, we do not treat the issue of detecting mentions, known as Named Entity Recognition, or of classifying them, namely, Named Entity Classification, as these problems have been tackled since quite some time by the NLP research community and several solutions exist nowadays[2]. Here, we focus instead on a relatively newer problem which concerns finding candidate referents for each mention in a Knowledge Base (KB) typically available as Linked Data, and choosing the right one by adding the corresponding link to the mention itself. In other words, given the TEI input text (in French)

le philosophe <persName>Voltaire</persName>

and the reference base DBpedia, we want to be able to automatically produce the output

le philosophe <persName ref="http://dbpedia.org/resource/Voltaire"> Voltaire </persName>

This task is commonly referred to as Named Entity Linking (NEL). NEL actually accomplishes two tasks at the same time, not only enrichment but also disambiguation. In fact, an entity is usually mentioned in the text in ambiguous forms. For instance, to remain in the literary domain, the mention "Goncourt" can refer to any of the two Goncourt brothers and writers, Edmond or Jules. At the same time Jules de Goncourt can be referred to in the text as "Goncourt", "J. Goncourt", "J. de Goncourt", etc. Besides, in order to automatically retrieve all passages in a text where Jules de Goncourt is mentioned, it is necessary not only to annotate all these mentions as named entities (NE) of the class person, but to provide them with a unique key that distinguishes them from those of other people, in this case those of Edmond de Goncourt. TEI annotation of named entities allows for different types of keys, in this case, for instance, we may use the bibliographic identifier "Goncourt, Jules de (1830–1870)", as well as the link <http://www.idref.fr/027835995>, pointing to the French identity reference catalogue entry for Jules de Goncourt. Proper linking can only be achieved by choosing the second strategy, and adding an external link to each mention. Ideally, the link should also point to a source containing additional and machine readable information on this author (e.g., birth date, birth place, authored works, etc.).

Linked Data (LD) [4] is a standardised way of publishing knowledge in the Semantic Web and many of the available data sets are of great interest for the DH [3]. LD principles such as interlinking and vocabulary reuse facilitate manipulation of data from heterogeneous sources. The formalised knowledge published in the form of LD can provide the background information required to disambiguate NEs in a given context by means of reasoning. In addition, such external knowledge enriches the text by remaining available in the annotation as a reference, and can be accessed at later stages for further processing; for instance, further connections can be found for the links by using information discovery systems such as in [5]. The increasing volume and availability of Linked Data brings new opportunities to build LD-based tools. In this context, LD sets serve

[2] It is noteworthy to mention that these solutions does not seem to be well-adapted to the DH (see paper [3] for a review of the difficulties).

frequently as external KBs to NLP tasks such as NEL[3]. The quality of the LD sets used as reference bases – in particular the completeness in terms of entities and the richness of relations between them – is crucial in NEL (see [7] for a discussion of LD quality and the formalisation of metrics adapted to NEL). Indeed, better datasets can help the algorithm in choosing the right referent, at the same time, they will better serve the final purpose of discovering and connecting information in the annotated texts.

In order to illustrate this latter point, we chose to evaluate NEL on a quite peculiar task (with respect to current research), namely, that of NE linking of authors in corpora of French literary criticism and essays from the 19th–20th centuries[4]. Such texts contain mentions of well known authors, such as "Hugo" and "Zola", but also lesser known critics such as "Barre"; the "right" reference base for this text will describe the entities Victor Hugo, Emile Zola and André Barre, identified by means of Uniform Resource Identifiers (URI), and will also include assertions concerning a series of shared relations and concepts, such as the fact that the first two were French writers, novelists, lived in the 19th century, etc. while the latter was the author of an essay on symbolism.

In this paper we shall first present previous approaches to NE disambiguation and linking, then introduce our graph based disambiguation algorithm, named REDEN, which includes strategies to consistently handle multiple LD sets, showing how this and other design characteristics make it better suited to work with texts from the humanities and literary domain. REDEN has been introduced in previous publications[5] [9], [8], [10]. Here, in contrast to our previous work, we shall extensively describe the algorithm in its two phases, namely, candidate retrieval and candidate selection, along with new improvements such as the exploitation of any KB linked to an entity using equivalence links (e.g., sameAs). We then define evaluation measures for the assessment of each of them both separately and in conjunction, and subsequently describe new experiments carried out along with their results. Finally, we shall draw some conclusions and discuss the lessons learnt especially in the light of the development of a NEL tool suited for the domain of DH.

2 Related Work

In this section, we define NEL according to current state of the art, subsequently we focus on the review of graph-based NEL approaches and, finally, we highlight the particular issues of domain adaptation for the use of NEL in the Digital Humanities.

2.1 NEL: Task Definition

We define NEL as the task of finding the referent to a NE mention in an input text, choosing between potentially different candidate entries in a knowledge base, and annotating the mention with the URI of the correct entry, if it exists in the KB.

NEL belongs to a family of related NLP and Information Retrieval tasks, having similar but not identical purposes. We do not intend to provide here an exhaustive survey of such tasks (see, for instance, [11], [12]), but we briefly clarify the problem definition, with respect to related approaches. In our definition, NEL is a more specific type of Named Entity Disambiguation (NED), given that disambiguation *per se* does not imply the identification of the referent for each entity, whereas linking implies disambiguating. NEL, instead, is more similar to the so called Wikification task, where each entity is linked to the relevant Wikipedia article. In particular, NEL is similar to what the authors of [11] refer to as Disambiguate to Wikipedia (D2W), with a Text and a Set of mentions as input and a Set of relevant annotations (links to Wikipedia) for each mention as output.

[3] For the connections between linguistics, NLP and linked data see [6].

[4] http://obvil.paris-sorbonne.fr/corpus/critique/

[5] The present paper is an extended version of [8], also a peer-reviewed version.

In our case though, the reference base is rather a linked data set, such as DBpedia or other domain specific ones, as we shall later see. A similar task to D2W and NEL is Word Sense Disambiguation (WSD) more generally, namely, the task of identifying the sense of polysemous words. A crucial difference is that in WSD we assume that the KB used for disambiguation, often a computational lexicon that lists senses for words, is complete, whereas in D2W, as in NEL, this assumption does not hold [13]; generally NEL algorithms should assign a null link to entities without a referent in the KB. For this reason, NEL will allow for null links (NIL) for the cases when the correct referent of a NE mention is not present in the KB.

Having said this, many authors use the aforementioned terms in a different way. For instance, [14] defines NED as the task of linking entity mentions in a text to a KB whereas they reserve NEL for the complete task of discovering (complete or potentially partial) mentions of entities within a text and to link them to the most suitable entry in a reference KB. We prefer to consider Named Entity Recognition and Classification (NERC) as a logically separate task from the disambiguation and linking, though many tools perform all of them together. Also, we emphasize the task of linking over that of disambiguating, since the semantic enrichment of texts is the crucial goal of our endeavour.

While being distinguished from NERC, the NEL algorithms also generally perform two logically separated steps: (1) retrieval of the candidates from some pre-processed source having an index or dictionary structure and (2) identification of the correct candidate. The first step is performed by means of string matching; indices are built, containing surface forms as keys, each associated with the links to all possible referents. Different types of search strategies can be used to improve retrieval such as the use of edit distance for string matching as well as the expansion of surface forms. For instance, the name “Emile Zola” can be expanded to generate three surface forms (“Emile Zola”, “Zola”, “E. Zola”). As for the second step, NEL algorithms can be coarsely divided in two different groups: those using text similarity and those using graph based methods for ranking the candidates and select the best one. Both these methods are unsupervised, and they do not rely on pre-annotated corpora for training.

The best known tool of the first group is DBpedia Spotlight [15], which performs NER and DBpedia linking at the same time. Spotlight identifies the candidates for each mention by performing string similarity between the mention and the DBpedia labels, then it decides which entry is the most likely the sought one by comparing the text surrounding the mention with the textual description of each candidate. The referent whose description is more similar to the context of the mention in terms of TF/IDF is chosen. This method is known to be very efficient, but it can only provide linking towards resources such as DBpedia, whose entries come with a description in the form of unstructured text. Other knowledge bases do not provide a textual description for their entries, such is the case of the bibliographical databases that constitute the ideal linking for mentions of authors. In this paper, we do not focus on text-based NEL tools, instead we review graph-based NEL which is the focus of the next subsection.

2.2 Graph-Based Approaches to NEL

Graph-based approaches to NEL are unsupervised algorithms relying on existing knowledge bases (e.g., the Wikipedia article network, Freebase, DBpedia, etc.). Reasoning can be performed through graph analysis operations. It is thereby possible to at least partially reproduce the actual decision process with which humans disambiguate mentions. In particular, these approaches build a graph out of the candidates available for each possible referent in a given context then use the relative position of each referent within the graph to choose the correct referent for each mention. The

graph is built for a context, such as an entire or a portion of text, containing possibly more than one mention[6], so that the disambiguation of one mention is helped by the other ones.

Several tools such as AIDA or NERSO [16], [17] and others mentioned hereafter, use graph based approaches for candidate disambiguation, as they seem quite promising. Generally speaking, such algorithms are based on the notion of node prominence in graph theory to identify the most likely candidate given a graph of candidates and their relations. Several approaches exist to compute prominence, the most relevant ones are centrality scoring (node degree, PageRnk) [18], semantic relatedness between candidates by random walks on graphs [14] or a voting system (TagMe) [19], best joint mention-entity mapping [20], graph-distance minimisation between candidates [21], etc. It is noteworthy to mention that these approaches deal solely with Wikipedia and more recently DBpedia as KB.

In this paper, we concentrate on centrality measures following similar approaches in Word Sense Disambiguation [22]. Centrality measures may be performed on the KB structure in order to use the rich set of relations to disambiguate mentions. For instance, in [23] English texts were disambiguated using a graph that relied only on English Wikipedia, and was constituted of the links and of the categories found in Wikipedia articles. For instance, the edges of the graph represent whether ArticleA links to ArticleB or whether ArticleA has CategoryC. Centrality is then used to assign the correct link to the ambiguous mention. Centrality is an abstract concept, and it can be calculated by using different algorithms[7]. In [22], the experiment was carried out using the following algorithms: *Indegree*, *Betweenness*, *Closeness*, *PageRank*, as well as with a combination of all these metrics using a voting system. Results showed the advantage of using centrality with respect to other prominence measures.

As for the KBs, generalistic ones such as DBpedia or Yago are the most cited ones, whereas experiments with domain specific KBs are less frequent. To this respect, [18] rightly insist that a NEL algorithm should ideally be agnostic as to the type of KB used, but in the most of tools, such as DBSL, it is very difficult to replace DBpedia with a custom or domain specific KB. Indeed, the authors developed their own tool, AGDISTIS, that allows for the use of any linked data resource provided with a SPARQL endpoint. However, this tool, as the aforementioned ones, does not take advantage of the main strength of Linked Data, which is the possibility to access more LD sets available through equivalence links and thereby to enrich the graph of candidates.

To this respect, a special mention should be given to Babelfy [14] a graph based tool that performs WSD and NEL at the same time. Given the peculiarity of this approach, it can only be performed with BabelNet, a specially designed KB built by automatically linking various lexical databases with Wikipedia. Indeed, fusing different sources of information can greatly improve the richness of the graph, as well as the performance of the algorithm. A way to generalise this approach, at least for NEL, is to exploit equivalence links in the KB, pointing to other sources of information. In the next section, we shall see how our algorithm uses multiple LD sets as KB, by iteratively accessing and crawling the different LD sets available thanks to equivalence predicates (e.g., sameAs) and by applying the appropriate fusion strategies.

Evaluation of NEL should be carried out by analysing the performance of NEL algorithms independently from the NERC phase, which isn't always the case in literature. Indeed, many of the aforementioned tools and works do not provide any separate figures for NEL accuracy; most of tools do not even allow users to input texts with pre-detected mentions to links. This makes comparisons of NEL performances alone rather difficult. As to the linking performance, their definition vary according to the task definition; for instance, in NERC + NEL tools, precision, recall and F-MEASURE follow the classical information retrieval definition. In other cases [11],

[6] TEI encoding makes the structure of the document machine readable, by allowing for the explicit markup of textual subdivisions such as sentences, paragraphs, chapters, parts. In this sense it makes the choice of the disambiguation context more straightforward.

[7] For a discussion of the notion of centrality see also [24].

authors are also interested in measuring whether the algorithm is consistently precise over a single and also a large number of documents, distinguishing macro- and micro-accuracy, respectively.

2.3 NEL: Domain Adaptation for the Digital Humanities

While the aforementioned approaches have been mainly developed and evaluated on news texts and web documents, there has not been – to the best of our knowledge – much research on NEL domain adaptation for the humanities and the literary domain in particular. [25] is one of the first works to highlight the importance of NEL for humanities, with a focus on toponyms. It also contains interesting reflections on the problem of temporal information, and how important it is to exclude some candidates (the correct referent must be a place that exists at the time in question); we shall later see how REDEN also exploits this intuition. Other papers concern the adaptation of Wikification approaches to the domain of cultural heritage, not limiting themselves to the enrichment of texts but also of digital metadata records associated to cultural artifacts. [26] adapts a known algorithm (WikiMiner), using Wikipedia to select correct links. Domain adaptation is performed via category pre-selection (culture, arts, humanities). [27] and [28] follow a similar approach.

Keeping in mind the aforementioned approaches and the limitations for their use in the DH, we have developed a graph-based NEL approach which is well-suited to handle the semantic annotation of digital literary editions and more broadly other domains in the DH. The algorithm is described in the next section along with our motivations and the user requirements.

3 Domain-Specific Graph-Based NEL Approach

The algorithm we are going to present was developed in the context of Labex OBVIL[8], an interdisciplinary French laboratory where computational methods are developed and applied to the research in the literary domain. OBVIL has developed a large corpus of TEI digital editions, both of primary and secondary literary sources and essays more generally. OBVIL stands for *Observatoire de la Vie Litéraire*, and literary life is investigated in its broader sense, with its ramifications and intersections with the cultural, scientific, and artistic aspects in each epoch; an example is the CORPUS CRITIQUE[9], a large diachronic collection of French essays. Following the current trends in digital literary studies [29], [30] tools are required to enrich texts for knowledge discovery and visualisation.

In particular, the linking of already tagged mentions is considered as a particularly painstaking operation for annotators, especially for persons names, as it requires the verification of several interlinked sources to identify the correct referent and the identification of the correct IDREF[10]. The analysis of available algorithms and tools for linking soon revealed their limitations, in terms of supported input – no TEI support, difficult to work with pre-detected mentions – but also in terms of supported document bases; in particular manual annotators observed that, in essays from the 19th and early 20th centuries, minor authors and members of the cultural life in general that did not have a Wikipedia (and thus DBpedia entry) could be identified by searching on data.bnf.fr, the linked data catalog of the *Bibliothèque Nationale de France (BnF)*, that is publicly available. Moreover, they observed that often the right referent was found by looking at the context of other mentions, using the relations between individuals, and combining information from different

[8] http://obvil.paris-sorbonne.fr

[9] http://obvil.paris-sorbonne.fr/corpus/critique

[10] IDREF - http://www.idref.fr - is the French reference base for individuals, also used in library catalogs; IDREF is an almost complete reference base for the domain in question, but contains very little contextual information on people and is not provided with a SPARQL endpoint; at the same time, many other LD sets are connected to it.

sources. Finally, they noticed that some candidates could be excluded *a priori*, since they were not even born at the time when the essay in question was written.

To recapitulate, the algorithm must fulfill the user requirements by being:

- TEI compliant
- LD based, it should access any user selected KBs (domain specific but also generalistic ones) and should also be available via an Sparql end point so as to be always linking to latest versions
- independent from NERC and, thus, able to deal with already tagged texts and be used with different NERC tools along with distinct manually corrected data
- adaptable to a user-defined domain scope
- able to deal with one class or sub class of NEs at a time

We, thus, decided that an adapted NEL algorithm should fulfill such requirements, firstly being TEI compliant. Indeed, REDEN adheres to currently recommended TEI formalism for the annotation of places and people in texts. At the same time, we are aware of current proposals for more complex annotation schemes within the TEI community, see, for instance, the GEOLAT project [31], where the relationships between entities are made explicit in text, while at the same time creating more adequate and domain specific data sets.

As mentioned beforehand, another requirement is that REDEN should be able to support potentially any LD set with a SPARQL endpoint, capable of gathering and fusing further information from new LD sets when equivalence links are present, and with a customisable index building facility that allows for an ad hoc creation of aliases and for the definition of specific time and space constraints.

Additionally, the proposed NEL algorithm should take a perfectly annotated text in TEI as input, i.e., entities are properly tagged and classified; such algorithm, we believe, would better integrate in an editorial work-flow of digital editions, where automatically detected NEs are first manually corrected, then automatically linked and then manually checked again. Moreover, NEL dissociated from NERC allows for thorough evaluation of the linking performance in an independent way.

Lastly, a graph based approach was chosen, to mimic the same disambiguation process that the annotators themselves used, as we shall see in Subsection 3.1. Furthermore, these kinds of approaches seem to be more easily adaptable to various domains and language independent. We named our algorithm REDEN, that stands for *Referencement et Desambiguisation d'Entités Nommées*, the French for Disambiguation and Referencing of Named Entities. As most NEL algorithms, REDEN performs two phases, namely, candidate retrieval and candidate selection. From a more technical perspective, it is important to distinguish both phases and evaluate them separately, following [13]. By doing this, we shall first check whether the candidate retrieval algorithm is able to produce sets containing (among others) the correct referent, and then we shall evaluate whether the chosen centrality measure ranks the correct referent higher than its wrong competitors.

In what follows, we shall first illustrate how REDEN works by an example on authors linking and exhaustively describe the proposed NEL approach.

3.1 General Intuition and Illustrative Example

Let us consider the following paragraph excerpt of a French digital edition of literary criticism entitled "Réflexions sur la littérature" (Reflections on Literature) written by Albert Thibaudet (1874–1936) and published in 1936[11]:

Mikhaël *et* ***Samain*** *se rapprochent du Parnasse et de* ***Baudelaire*** *bien plus que de* ***Verlaine****. C'est voir* ***Jammes*** *par un très petit côté, qu'en faire un "excentrique", c'est abuser de certains*

[11] The TEI edition was published in 2014 by Labex OBVIL and can be found online, http://obvil.paris-sorbonne.fr/corpus/critique/thibaudet_reflexions/.

excès voulus, et en somme le petit veau qui était pauvre, ou la vache qui a mangé les bas noirs de la fiancée du poète, sont-ils plus "excentriques" que bien des ballades de ***Laforgue****?*

In bold, we see six mentions that were properly recognised by a NER algorithm, that now have to be linked to an URI. For each mention, REDEN selects the URIs of the candidates from a customised domain-adapted index of author surface forms that is automatically built beforehand out of a reference linked data set, the most representative one of the domain in consideration (e.g., BnF). The constitution of the index along with the strategies to search for candidates within it represents the first phase of NEL, that is, **candidate retrieval**.

An excerpt of the resulting candidates for the six NE mentions (along with the number of candidates) from the example is shown below[12].

Candidates **(1)** (**Mikhaël**) = Éphraïm Mikhaël (1866–1890)
Candidates **(1)** (**Samain**) = Albert Samain (1858–1900)
Candidates **(2)** (**Baudelaire**) = Charles Baudelaire (1821–1867), Auguste Colas dit Baudelaire (1830–1880), ...
Candidates **(5)** (**Verlaine**) = Paul Verlaine (1844–1896), Madame Paul Verlaine (1853–1922), ...
Candidates **(4)** (**Jammes**) = Francis Jammes (1868–1938), Geneviève Jammes (1882–1963), ...
Candidates **(3)** (**Laforgue**) = Jules Laforgue (1860–1887), René Laforgue (1894–1962), ...

Per mention, REDEN subsequently retrieves the RDF graphs of the resources corresponding to the candidate URIs from the reference LD set (e.g., BnF). At the same time, REDEN uses existing equivalence links (e.g., *sameAs*) in order to retrieve also the RDF graphs of their homologous resources described in other LD sets (e.g., DBpedia, Wikidata, and any other available ones). Afterwards, graphs are combined into a single enriched graph by fusing the assertions of homologous resources. In other words, assertions about Jules Laforgue in the BnF data set and those from the same entity in the DBpedia data set are fused into one resource.

The combination of the aforementioned fused graphs for all mentions results into a larger graph where the vertices are URIs of the candidates and of other kinds of resources or literals and edges represent RDF predicates; optionally, weights can be assigned to specific predicates. REDEN prunes the resulting graph so that it contains only those edges involving at least two candidates of different mentions, in order to retain only the predicates that would play an important role in the disambiguation process. Once the combined candidates graph is ready, the calculation of centrality (e.g., Degree or Eigenvector centrality) is performed for each candidate and is then used to choose for each of the six mentions the correct (= most central) referent. The proper constitution of the fused graph and the graph-centrality calculation constitute the second phase of the NEL algorithm, namely **candidate selection**.

Figure 1 shows an excerpt of the resulting graph where candidates are underlined and the best candidates per mention (chosen by the algorithm) are marked in bold. For readability sake, only intermediary nodes that most influenced the choice of the best candidate per mention are shown in the figure. Notice that edges in RDF graphs are typically directed; as the centrality algorithms we use do not take into consideration edge directionality, we decide not to display edge directions.

Generally, edges of mention candidates tend to very generic categories; in this example, for instance, we omitted to show common *rdf:type* edges to the vertex of the category *Human*, *Person*, and *Male* as the majority of candidates are men; these categories are formally described in the DBpedia or Wikidata ontologies. Some of the chosen candidates also share *rdf:type* an edge to more specific categories such as *Writer Artist* or *yago:Poet110444194*, which belong to the so-called Yago categories derived by [32] from the Wikipedia category model. But, as we can observe from the figure, the vertices *FrenchPoets* and *SymbolistPoets* are those that influence the final choice the most, giving the correct candidates a higher centrality over their competitors. Indeed, the passage above refers to French poets, being Charles Baudelaire and Paul Verlaine

[12] In this example, we identify candidates by distinguishable personal information instead of URI for readability sake.

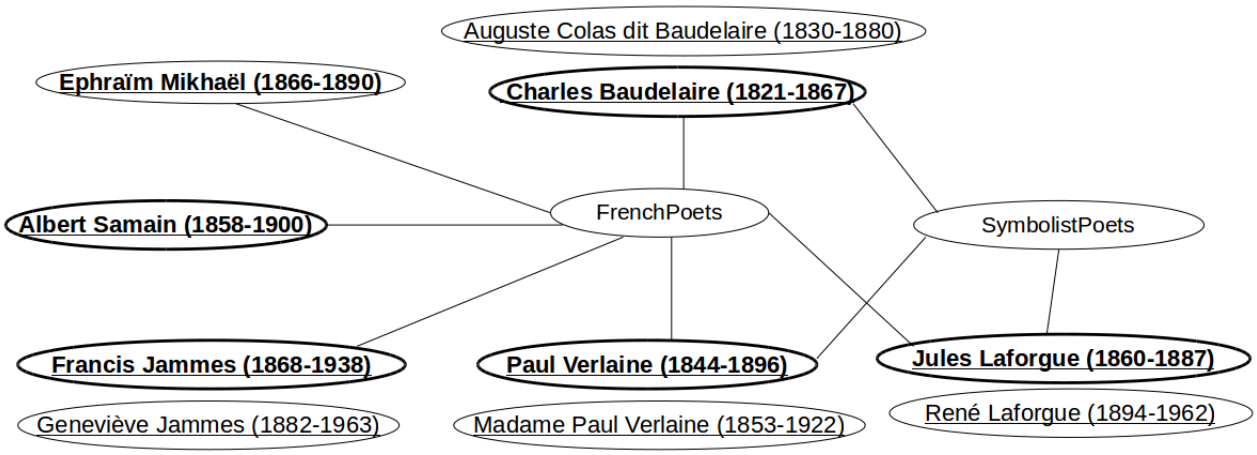


Figure 1. Excerpt of the chosen URIs (in bold) for the six candidates (underlined); here, all edges represent *rdf:type* links

the most notable figures of the epoch. They both belonged to the Symbolism movement, just as Jules Laforgue. In other cases, it is possible to retrieve links such as *influencedBy* or *influences* that can sometimes play an important role for choosing the correct candidate. It is interesting to notice that even experts in French literature may know little about the life and work of some of the minor poets mentioned in the text; here in particular, they were not entirely certain of the identity of "Samain" and "Mikhaël", but REDEN's choice of Albert Samain and Éphraïm Mikhaël was then judged to be the correct one. Indeed, the automatic discovery of the identity of these less-known authors facilitates the work of experts to a great extent.

3.2 Description of the Algorithm

With REDEN[13], we propose a graph-based, centrality-based approach. The algorithm processes a XML-TEI file where NE mentions are already tagged (e.g., <persName>, <placeName>, <orgName>) and outputs an enriched version of the input file where URIs are assigned to these mentions. Figure 2 presents a flowchart to illustrate the REDEN algorithm.

Alternatively, the pseudo-code of the core algorithm is presented in Algorithm 1. While most of it refers to the candidate selection phase, only lines 1 and 3 refer to the candidate retrieval phase; notice that the constitution of domain-adapted indices is briefly described later on. In the following subsections, we further describe each phase and discuss the strategies we chose for properly handling linking of authors in French literary essays.

3.2.1 Candidate Retrieval

As previously stated, REDEN searches for the tagged mentions within a text portion of the input TEI, henceforth called context (e.g., paragraph, chapter or whole text). REDEN adopts the one-sense (or, in this case, referent) per discourse approach [33] within a chosen context, as typical in WSD algorithms. Conveniently, mentions may be tagged using the different possibilities offered by the XML-TEI standard[14]; our algorithm is, thus, able to use XPath expressions to provide more flexibility concerning the choice of mentions to be processed in the TEI; for instance, the following expression, $persName[not(@type =' character')]$, searches for all mentions tagged as persons except those being referred to as fictional characters.

[13] Code source and useful resources can be found here: https://github.com/cvbrandoe/REDEN.
[14] http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ND.html

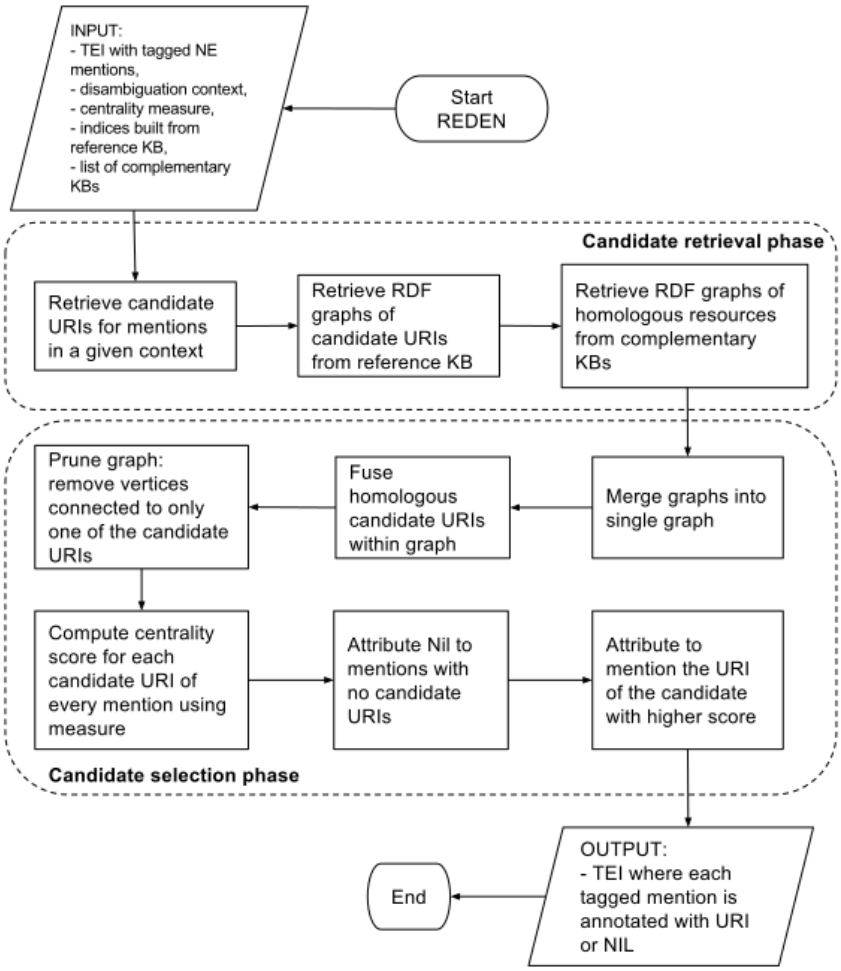


Figure 2. Flowchart to illustrate REDEN algorithm

Subsequently, for every mention, REDEN searches for candidate URIs by exact string matching in an index of surface forms per NE class (e.g., authors)[15]. Such an index is automatically built, only once, out of data from a selected reference LD set. The choice of the reference LD set is crucial for proper retrieval, as it needs to comply with some quality requirements discussed in the following (also see in [7] the procedure to evaluate *a priori* the quality of LD for NEL). The LD set must be the most representative of the domain at stake in terms of completeness. Regarding authors mentioned in French literary essays, while some of the most famous authors have rich entries in broad-coverage ontologies such as DBpedia, other less known ones are only present in domain-specific KBs such as BnF. For this reason, the choice of BnF as the reference source seems the most convenient because it is the most complete source for our purposes [9].

Another necessary condition for proper retrieval is the existence of interlinking to other LD sets; for what concerns our choice: BnF links to DBpedia, thus, making it very easy to further retrieve and combine more information in one knowledge graph. Besides DBpedia, bibliographic datasets like BnF also strongly rely on domain specific standards such as idref, viaf, or ISNI for interlinking. Sometimes interlinking may follow different strategies, so, for instance, sources such as DBpedia and Yago usually make use of *sameAs* relationships; but other LD sets like BnF use other relations with the same semantics, such as *skos:exactMatch*.

[15] In some cases string similarity algorithms may overcome minor spelling variations, but not major ones, which require proper information on NE alternative naming forms.

Algorithm 1 Simplified pseudo-code of the proposed NEL algorithm.

Require: mentions: list of mentions, measure: the centrality measure,
context: the size of the disambiguation context
build only once domain-adapted indices
for mention in mentions for a given context **do**
  candidate URIs ← retrieve candidate URIs for *mention*
  graph ← retrieve RDF graphs of *candidate URIs*
  multi-source graph ← retrieve RDF graphs of the homologous resources of *candidate URIs*
  merge *graph* into *multi-source graph*
  fuse homologous *candidates URIs* within *multi-source graph*
end for
prune *multi-source graph* by removing vertices connected to only one of the candidate URIs
for mention in mentions **do**
  if *candidate URIs* is empty **then**
    mention URI ← NIL
  else
    score ← compute centrality for each *candidate URI* using *measure*
    mention URI ← choose the candidate URI with higher centrality *score* for *mention*
  end if
  annotate *mention* with *mention URI*
end for
return each mention annotated with URI or NIL

Moreover, either the reference LD set or at least one of the interlinked ones must exhaustively describe entities in terms of relations with other entities and of alternative labeling properties (e.g., *rdfs:label*, *skos:altLabel*). For our purposes, DBpedia and BnF seems to be the most relevant sources. DBpedia describes authors reusing widely-accepted vocabularies (e.g., foaf, skos); authors are linked to each other by semantic relations such as *influencedBy*, and, indirectly, by being linked to the same concept, such as *SymbolistPoets*. BnF entries list all authors of books ever published in France; their entries contain information on name, date of birth and death, gender, authored works. For instance, the BnF entry for Voltaire[16] gives several alternate names such as François-Marie Arouet (Voltaire's real name), Wolter, Good Natur'd Wellwisher.

In order to automatically build the index out of the reference LD set, we built a domain-adapted LD extractor [10] included in REDEN. It performs SPARQL queries and enables the definition of temporal and spatial extents for constraining data; Figures 3 and 4 illustrate how it functions. Such queries use widely-accepted properties such as names (*foaf:name*, *foaf:familyName*, *skos:altLabel*) for filtering and retrieving exact matches from mentions. Other complementary properties such as foaf:gender help to match mentions containing honorific titles (e.g., M. Vigny, Madame De Staęl, etc.). To reduce the waiting time of query response retrieval, the local index per class (e.g., Person) is built and updated regularly. The resulting index lists automatically generated forms and their associated URIs such as: surname only (Rousseau), initials + surname (J.J. Rousseau, JJ Rousseau, ...), title + surname (M. Rousseau, M Rousseau), etc. This procedure ensures the retrieval of at least one candidate URI for most mentions. At the same time, the mass of information present in the BnF repository will generate several homonyms and make most mentions ambiguous, thus, good disambiguation becomes crucial. For other types of entities such as places, the procedure to build the index is straightforward using labeling properties such as *skos:altLabel* or *rdfs:label*; we have already built place indices from GeoNames and DBpedia for the experiments in [34].

[16] http://data.bnf.fr/11928669/voltaire/

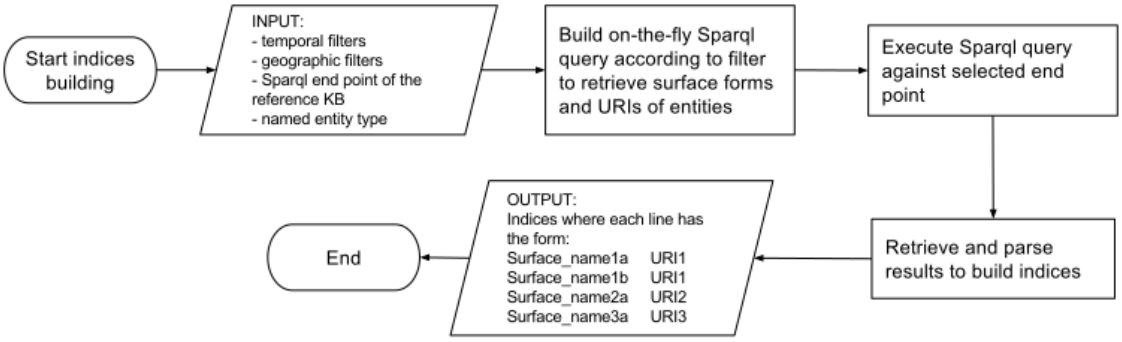


Figure 3. Flowchart describing the domain-adapted data extraction performed by REDEN during candidate retrieval

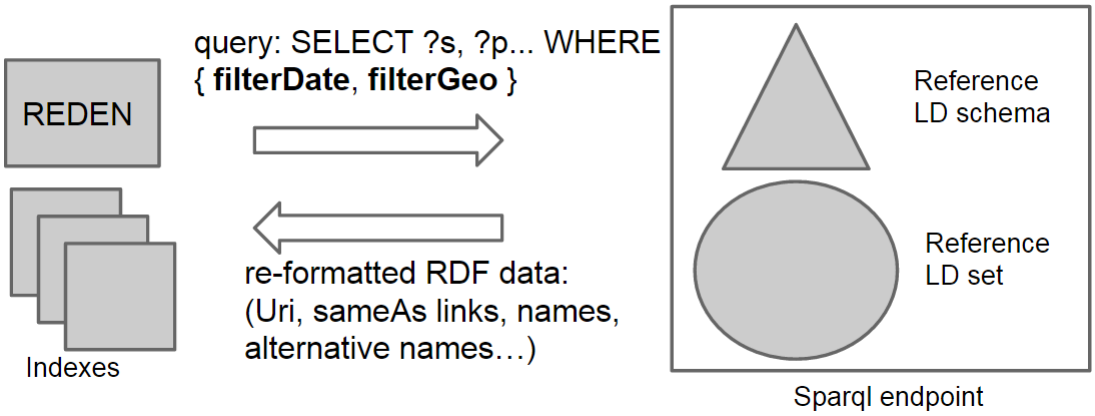


Figure 4. Domain-adapted data extraction performed by REDEN during candidate retrieval

3.2.2 Candidate Selection

As previously anticipated, our algorithm requires the construction of the relevant graph for the selection of the best candidates. This graph must represent domain knowledge while avoiding at best, redundant and conflicting information; it also must possess relevant knowledge for the disambiguation process. In the presence of multiple LD sources, it is, thus, important to properly fuse assertions of different LD resources describing the same entity into a single reference resource (e.g., BnF). Clearly, fusion may be performed with strategies of great complexity, such as the ones commonly used for publishing linked data and implemented in tools such as in [35]. At the moment, there is no need for this level of complexity for our purposes.

Our fusion process can be described in more detail as follows. Take a mention M (e.g., Hugo), which corresponds to a real World entity E and has the candidates C_1, C_2, ... (e.g., Victor Hugo, François Victor Hugo). Each candidate possesses a set of URIs from several LD sources (e.g., Victor Hugo from BnF, Victor Hugo from DBpedia, and so on). In particular, let us name the LD sources as LD_{ref}, LD_1, D_2, ... where LD_{ref} is chosen to be the reference source. It is straightforward to obtain the corresponding RDF graphs and convert them in equivalent undirected and unweighted graphs [36] named G_ref={V_{ref}, EG_{ref}}, G_1={V_1, EG_1}, G_2={V_2, EG_2}, ... Vertices V represent URIs of mention candidates of the reference LD set and asserted ontology concepts (e.g., *dbpedia:Writer*) and data-typed literals (*bio:birth*). Edges EG designate binary and labeled relations (e.g., *rdf:type*).

The aforementioned entity E is represented as E_{ref} in graph G_{ref}, as E_1 in G_1, E_2 in G_2, ... There exists an equivalence link (e.g., *sameAs*, *skos:exactMatch*) defined among E_ref and its

equivalent E_1, E_2, ... In this manner, the iterative fusion of G_{ref}, G_1, G_2, ... results in a graph G_f where E_{ref} identifies the entity E which is the product of the fusion. E_{ref} inherits the set of edges along with the corresponding vertices in which objects of the relation (in RDF terminology) are their homologous in the other sources. It is also possible to *a posteriori* assign weights to these edges based on user preferences, in particular, the higher weight is attributed to an edge, the more priority is given to this edge during centrality calculation. In [8], we describe an attempt to evaluate the impact of relations by setting a higher weights on them.

Once the fusion is completed, irrelevant edges are removed from the graph: only edges which involve at least two vertices representing candidate URIs are preserved in the graph. Finally, the proper centrality algorithm can be applied as parametrised in user preferences; currently implemented measures are: *DegreeCentrality*, i.e., the number of in and out links of a node; *BrandesBetweennessCentrality*, i.e., the number of times a node acts as a bridge along the shortest path between two other nodes; *FreemanClosenessCentrality*, i.e., the length of the average shortest path between a node and all vertices; *EigenVectorCentrality* which is based on the principle that a node is important if it is linked to by another important node. These measures are typically used in social network analysis and the word sense disambiguation, and rely on the implementation offered by the JgraphT-SNA library[17]. The selected centrality measure is applied to all candidates of each mention, and the best connected candidates (scoring higher with respect to their competitors) are chosen as referents; an enriched version of the input TEI file is then produced, by adding the URI of each mention. In its default settings, the system adds only the URI from a specified reference base, but it can also display all of the equivalent URIs for that candidate that was retrieved from the other connected sources. Finally, Figure 5 illustrates once more the entire workflow of the algorithm, this time using an excerpt of the example presented in Subsection 3.1.

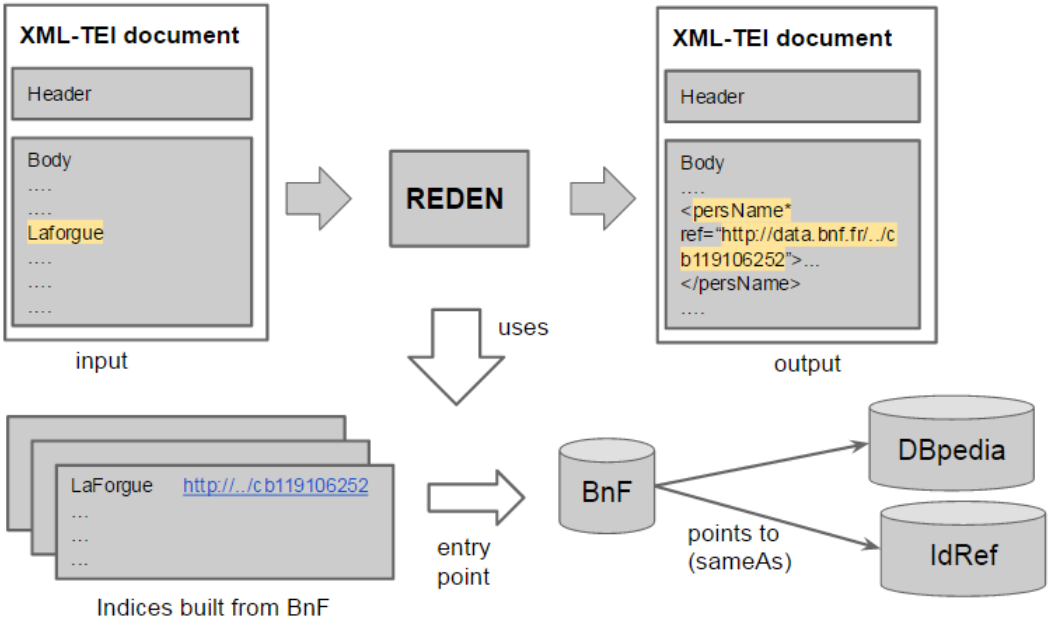


Figure 5. REDEN general work-flow adapted to the example in Subsection 3.1

[17] https://bitbucket.org/sorend/jgrapht-sna

4 Experiments and Results

This section describes the experimental settings used to test our proposal along with our findings.

4.1 The Test Corpora

As anticipated above, we use a perfectly annotated source, that is where entities are identified and classified, for which manually checked links are also available. We concentrate on the class *Person*.

The test corpora consists of two texts from the aforementioned CORPUS CRITIQUE, a French text of literary criticism entitled "Réfléxions sur la littérature" (Reflections on Literature) published by Albert Thibaudet in 1936, and a scientific essay entitled "L'évolution créatrice" (Creative Evolution) written by Henri Bergson and published in 1907. Both texts are quite rich in NE mentions of individuals, particularly authors, scientists, artists, but are different in style and in the density of references. Mentions of persons in these text were manually annotated by experts[18]; URIs assigned to mentions are those from IDREF, or NIL when experts did not know to whom the mention refers to or could not find an entry in IDREF.

The resulting test corpora contains 2980 (Thibaudet) and 380 (Bergson) tagged mentions of person entities where 1911 and 277, respectively, are manually annotated.

4.2 Evaluation Measures

We provide evaluation measures that allow us to analyse the performances of both phases of REDEN, candidate retrieval and candidate selection. In particular, we are interested in checking how good is phase one in retrieving the correct candidate, how good is the centrality based algorithm in choosing the right referent when more than one choice is present, and, finally, whether the algorithm is able to produce correct NIL annotations for those mentions that have no known referent in the gold.

Partly inspired by the work of [13], we define the following measures:

- Measures to Assess Phase One (Retrieval of Candidates From the KBs)
  - CANDIDATE PRECISION the proportion of non empty candidate sets containing the correct URI wrt the number of non empty candidate sets.
  - CANDIDATE RECALL the proportion of non empty candidate sets containing the correct URI wrt the number of all mentions that have a link in the gold.
  - NIL PRECISION the proportion of empty candidate sets for mentions that had NIL manual annotation wrt all empty candidate sets returned by phase one.
  - NIL RECALL the proportion of empty candidate sets that for mentions that had NIL manual annotation wrt the number of all mentions with NIL manual annotation.
- Measure Assessing Phase Two (Choice With Centrality Computation)
  - DISAMBIGUATION ACCURACY the proportion of correctly chosen links, when the candidate set contains the correct mention.
- Measure Assessing Overall Linking.

Finally, we also define an overall linking measure, that is intended to assess the goodness of the whole linking process.

- OVERALL ACCURACY the proportion of correctly linked mentions for the mentions that have a link in the gold.

[18] Fictional characters are also annotated in the text, but we considered that such entities do not strictly belong to the class of real individuals, and, thus, excluded them from the experiment. Linking of fictional characters is an interesting task, but requires a specific KB.

Measures for Phase One are particularly important to check whether the KB is fit for the purpose. In its current implementation, the REDEN algorithm will always choose a link when a non empty candidate set is returned[19]. Ideally, thus, the candidate retrieval phase should always return the correct candidate (among other possibilities) and possibly an empty set when no link exists in the KB. This of course is not always the case, due to homonyms and missing aliases. For this reason, it is interesting to evaluate Phase One in isolation, to see whether errors are due to wrong candidate selections, and to assess whether the graph based algorithm makes the correct choice when given the chance. The number of candidates per mention clearly is also an important measure, though not directly an evaluation measure, since it shows what level of ambiguity exists in the dataset. We, thus, calculate a further indicator:

– CANDIDATE CARDINALITY MEAN the average number of candidates per mention.

4.3 Experiment Settings and Results

For these experiments, we chose *DegreeCentrality* [37] as centrality measure because it has empirically proved in the previous work [9] to be the most satisfying one in our domain; evaluation was performed using standard correctness rates. In previous experiments [10], we also compared the correctness rates obtained by REDEN and a widespread NEL tool, DBSL. REDEN performed similarly to what state of the art graph-based NEL algorithms do in journalistic texts. The comparison was not totally fair for, as stated above, most algorithms do not allow for the separate evaluation of NERC and NEL.

Our algorithm allows for the user to customise the context of disambiguation, we chose the optimal mention context for each corpus, the *chapter* for Thibaudet and *whole text* for Bergson, as proved in [9]. Also, we do not assign weights to relations because, again as showed in [9], an adequate balance between mention density for the given context and a number of relations involved is generally enough for the algorithm to choose the best candidate for each mention.

Here, we concentrate on the thorough analysis of REDEN's performances with the aforementioned indicators, in order to assess the different parts of the algorithm, its weaknesses and its strengths. Table 1 presents the results of REDEN for both Thibaudet and Bergson using the aforementioned settings.

Table 1. REDEN evaluation results for the experiment

	Thibaudet	**Bergson**
Candidate precision	0.66	0.39
Candidate recall	0.95	0.46
NIL precision	0.80	0.94
NIL recall	0.17	0.48
Disambiguation accuracy	0.90	0.78
Overall linking accuracy	0.63	0.64
Candidate cardinality mean	4.56	3.94

4.4 Discussion Related to the Candidate Retrieval Phase

Concerning the first phase of the algorithm, as stated earlier, we want to check how good is this phase in retrieving the correct candidate. Candidate precision gives us the ratio of candidate sets per mention containing the right referent, here **0.66** for Thibaudet and **0.39** for Bergson. This

[19] As an alternative, a threshold for the centrality figure could be introduced, so that when it is too low no choice is made.

means that, in general but not systematically, REDEN finds the proper URI in the KB among non empty candidate sets; the exceptions concern those mentions missing manual annotations, in these cases, REDEN is able to retrieve candidates from the index but cannot know if the right one is in the candidate set. Manual annotation is missing most often because experts did not know or are not certain of the identity of the authors, or was simply an omission. An example of one of the aforementioned cases in Thibaudet concerns the mention "M. Clemenceau" which clearly for the expert refers to some unknown author of the epoch and not to the well-known French political figure. These cases are more common for Bergson than for Thibaudet, it is also related to the fact that the Thibaudet gold, in contrast to the Bergson gold, was subject to a more careful manual annotation process.

In complement to the previous measure, the candidate recall considers the cases when a manual annotation is present in the gold. Remarkably, an elevated ratio of **0.95** in Thibaudet points out that REDEN very frequently retrieves the right referent in the corresponding candidate sets when the entity is known by the experts and exists in the author index, in other words, the mention corresponds to an entity which can be accessed in the index by its real name or by any of its alternative names. In other cases, REDEN did not manage to find and retrieve the right one into the candidate set because the strategies for generating alternative naming for entities were not sufficient for matching the mention. Some minor spelling issues, e.g., Viélé-Griffin instead of Vielé-Griffin, are the source of these errors, though it does not have an important impact on the expected performances of the system. Also, some few cases concern entities having pseudonyms not listed in the KB, for instance, the alias William Stanley for William Shakespeare cannot be found in BnF. For Bergson, candidate recall is quite low, we can notice similar issues in larger amount; and also simply because the entry does not exist in the KB.

As mentioned beforehand, we aimed to assess whether the algorithm is able to produce correct NIL annotations for those mentions that have no known referent in the gold. Both NIL precision and recall can provide us with some useful hints. NIL precision which can be interpreted as: in the presence of NIL manual annotation, the proper referent does not exist in the KB, thus, the algorithm should confirm this fact and find no suitable candidate, in other words, the candidate set should be empty. Here, these measures are significantly high for both texts, **0.80** for Thibaudet and **0.94** for Bergson. However, in some cases, it seems that limited naming strategies can impact these measures by narrowing the access to the proper referent, for instance, few mentions marked as NIL such as "Della Rocca De Vergalo" do possess a referent in the KB but the naming strategies did not completely handle these particular cases. In other words, such errors are due to lacking or wrong encoding: individuals composite surnames such as John Stuart Mill, or Rémy De Gourmont seem to create more problems since the linked data sets do not provide the correct information on forename and surname.

On the other hand, NIL recall is considerably low, **0.17** for Thibaudet and **0.48** for Bergson, this implies that REDEN retrieves candidates from the KB even when the mention refers to someone that certainly does not have an entry in the KB. This is expected as the author index is quite large and has been automatically created, so it is expected to have many entries which correspond to a single entity. Therefore we count on a strong second NEL phase which will select the right candidate.

4.5 Discussion Related to the Candidate Selection Phase

Regarding REDEN second phase, here, we asses how good is the centrality based algorithm in choosing the right referent when more than one choice is present. The disambiguation accuracy results obtained seem very satisfying for both corpora, **0.90** for Thibaudet and **0.78** for Bergson, that is, the disambiguation process has enough contextual information for selecting the appropriate candidate; or, when the mention was annotated as NIL, it was due to the nonexistence of the right

entry in the KB. In some cases however, there was not enough information in the KB to let the algorithm to choose the right candidate.

In this matter, interestingly in Thibaudet we have 130 wrongly assigned mention instances, but they correspond to less than 20 distinct individuals, in Bergson wrong assignments are 14 for three distinct individuals; since our evaluation counts each instance of a mention as an error, it means that errors on frequent individuals can impact the evaluation to a large extent. For instance, the greatest source of error in Thibaudet is the mention "M. Barre" which refers to André Barre, a critic and an expert on symbolism, contemporary of Thibaudet, but is systematically mistaken for Joseph Barre, a professor of theology who lived in the 18th century. This case illustrates the problems issuing from the missing links in the used linked data bases; clearly, a connection with symbolism is visible by humans looking at André Barre's authored works, but is invisible to the machine in the form of explicit links to the categories relating to symbolist movements. In other cases, we can see that problems arise from the fact that REDEN cannot read temporal proximity in the same way as we do, since this is not often encoded in relations that can be read in the graph; so mentions of Payen are systematically assigned to Nicolas Payen (1512–1559) a musician, instead of the lesser known Fernand Payen (1872–1946) who was a contemporary of Thibaudet.

Mentions of the philosopher Plato are also subject to a systematic error (both in Thibaudet as well as in Bergson), and indeed quite an odd one: the system assigns to these mentions the URI of Vincenzo Cuoco (1770–1823), the Neapolitan philosopher, who authored a work in which he claims to be Platon traveling through Italy. BnF lists Platon among Cuoco's aliases, and so he is suggested as a candidate, and then due to some relations to other candidates in the context graph he gets promoted over the Greek Philosopher. This error teaches us two lessons: first, the information about aliases may be a source of noise, depending to the guidelines chosen by a given data set; second, that maybe some sort of prior knowledge should be incorporated in the algorithm in order to account for the fact that some mentions are strongly associated with a given referent independently from the context. In a number of cases though, the incorrect choice is quite plausible; so is, for instance, for the mention Nisard whose correct referent is Charles Nissard (1808–1889); the algorithm chooses his brother Desiré, who was also a critic and an academic, as well as a politician, and collaborated strongly with his brother.

Finally, the overall linking measure obtained, **0.63** for Thibaudet and **0.64** for Bergson, seems to summarise well the results of the other evaluation measures; in simpler terms, a good disambiguation accuracy affected though by a less efficient candidate retrieval.

5 Conclusions and Future Work

We presented an algorithm to perform NE disambiguation by referencing author mentions to broad-coverage and domain-specific Linked Data sets, DBpedia and BnF, respectively. We set up a procedure that extracts RDF data of person resources described in these LD sets. This procedure can be generalised to other classes of NE (e.g., places) only by modifying the corresponding SPARQL query. Furthermore, REDEN crawls RDF graphs from homologous resources described in other LD sets using equivalence links, and our fusion procedure enables for the constitution of a non-redundant graph which is well-suited for centrality calculation during the candidate selection phase. It is noteworthy to mention that REDEN is not completely language dependent; more specifically, the candidate selection phase is language-independent and the index constitution part during the candidate retrieval phase needs only the specification of the appropriate Sparql query to build an index in the desired language.

In the present paper, we performed experiments on French literary criticism texts and scientific essays from the 19th century and early 20th century with promising results. These findings will help us to study the weaknesses and strengths of our algorithm thereby to achieve further improvements. Crucially, we were able to obtain such good results while at the same time

developing an algorithm that follows current standards in digital editions (TEI) and semantic Web (RDF). This will facilitate the use of the algorithm by humanities researchers and allow for the use of various new data sets.

REDEN is an open-source[20] and we also offer resources to encourage its use by digital humanities and cultural heritage scholars. Ongoing developments are aimed to perfect and test REDEN in different contexts, so that it can build indexes from various data sets and efficiently disambiguate persons and other classes (notably places and organisation) in various domains. Sources such as Getty and Geonames have already been tested. In order to ensure usability, especially by digital humanists working on the enrichment of digital editions, a web based GUI for REDEN is currently under development. This tool, dubbed REDEN ONLINE [38][21], is currently customised for the linking of authors and places; additionally, it builds on-the-fly different kinds of visualisation for the input data by crawling available LD sources.

Further experiments will compare REDEN with other graph-based NEL approaches using a more significant amount of French Literature texts, which are being compiled and annotated. To compare with other systems, we shall evaluate REDEN in a scenario where entities are automatically detected and classified using existing NERC algorithms without manual checking before NEL is applied, in order to verify and measure the impact of NERC on NEL.

To conclude, the present work will certainly be of interest to those groups within the DH research community who are actively developing and experimenting with computational methods built in the context of NLP, Computational Linguistics, the Semantic Web and their different intersections, as well to the researchers in the aforementioned disciplines interested in domain adaptation and novel use cases. The adaptation of the proposed approach to other contexts is straightforward though it is conditioned by the availability and the richness of linked data and linguistic resources. Indeed, this particular experience has provided us with more knowledge about the difficulties of domain adaptation of tools and algorithms for the literary domain and more broadly the DH. At the same time, it has revealed the need of well-suited domain-specific LD-based knowledge bases, annotated corpora in languages other than English and rich linguistic resources providing homonyms for named-entities. Clearly the existence of useful applications relying on such resources – such as REDEN and REDEN ONLINE – could prompt the community to make more efforts in this direction.

Acknowledgements

This work was supported by French state funds managed by the ANR within the *Investissements d'Avenir programme* under reference ANR-11-IDEX-0004-02 and by an IFER Fernand Braudel Scholarship awarded by FMSH.

References

[1] T. Blanke and M. Hedges, "Scholarly Primitives: Building Institutional Infrastructure for Humanities E-Science," vol. 29, no. 2, pp. 654–661. Available: http://dx.doi.org/10.1016/j.future.2011.06.006

[2] L. Burnard, What is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources, ser. Encyclopédie numérique. OpenEdition Press. Available: http://books.openedition.org/oep/426

[3] S. V. Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. V. De Walle, "Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections," Literary and linguistic computing, 2013. Available: http://dx.doi.org/10.1093/llc/fqt067

[20] https://github.com/cvbrandoe/REDEN

[21] See http://obvil-dev.paris-sorbonne.fr/reden/RedenOnline/site/input-tei.html for beta version.

[4] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - the Story so far,” Int. J. Semantic Web Inf. Syst., vol. 5, no. 3, p. 1–22, 2009. Available: http://dx.doi.org/10.4018/jswis.2009081901

[5] B. De Meester, T. De Nies, L. De Vocht, R. Verborgh, E. Mannens, and R. V. de Walle, “Exposing Digital Content as Linked Data, and Linking Them Using StoryBlink,” in Proceedings of the 3th NLP&DBpedia workshop, Oct. 2015.

[6] C. Chiarcos, S. Nordhoff, and S. Hellmann, Eds., Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata. Springer, 2012. Available: http://dx.doi.org/10.1007/978-3-642-28249-2

[7] C. Brando, N. Abadie, and F. Frontini, “Linked Data Quality for Domain Specific Named Entity Linking,” in Proceedings of the 1st Atelier Qualité des Données du Web, 16ème Conférence Internationale Francophone sur l’Extraction et la Gestion de Connaissances, Reims, France, Jan. 2016.

[8] C. Brando, F. Frontini, and J.-G. Ganascia, “Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets,” in New Trends in Databases and Information Systems. Springer, 2015, pp. 505–514. Available: http://dx.doi.org/10.1007/978-3-319-23201-0_51

[9] F. Frontini, C. Brando, and J.-G. Ganascia, “Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts,” in Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, 2015, pp. 77–88. Available: http://ceur-ws.org/Vol-1364/

[10] F. Frontini, C. Brando, and J.-G. Ganascia, “Domain-Adapted Named-Entity Linker Using Linked Data,” in Workshop on NLP Applications: Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015), Passau, Germany, Jun. 2015. Available: https://hal.archives-ouvertes.fr/hal-01203356

[11] M. Cornolti, P. Ferragina, and M. Ciaramita, “A Framework for Benchmarking Entity-Annotation Systems,” in Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013, pp. 249–260. Available: http://dx.doi.org/10.1145/2488388.2488411

[12] A. Fader, S. Soderland, O. Etzioni, and T. Center, “Scaling Wikipedia-Based Named Entity Disambiguation to Arbitrary Web Text,” in Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA, 2009, pp. 21–26.

[13] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran, “Evaluating Entity Linking with Wikipedia,” Artificial intelligence, vol. 194, pp. 130–150, 2013. Available: http://dx.doi.org/10.1016/j.artint.2012.04.005

[14] A. Moro, A. Raganato, and R. Navigli, “Entity Linking Meets Word Sense Disambiguation: A Unified Approach,” Transactions of the Association for Computational Linguistics, vol. 2, pp. 231–244, 2014.

[15] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, “Dbpedia Spotlight: Shedding Light on the Web of Documents,” Proceedings of the 7th International Conference on Semantic Systems, pp. 1–8, 2011. Available: http://dx.doi.org/10.1145/2063518.2063519

[16] M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, “Aida: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables,” Proceedings of the VLDB Endowment, vol. 4, no. 12, pp. 1450–1453, 2011.

[17] S. Hakimov, S. A. Oto, and E. Dogdu, “Named Entity Recognition and Disambiguation Using Linked Data and Graph-Based Centrality Scoring,” in Proceedings of the 4th International Workshop on Semantic Web Information Management, ser. SWIM ’12. New York, NY, USA: ACM, 2012, pp. 4:1–4:7. Available: http://dx.doi.org/10.1145/2237867.2237871

[18] R. Usbeck, A.-C. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both, "AGDISTIS-Graph-Based Disambiguation of Named Entities Using Linked Data," in The Semantic Web–ISWC 2014. Springer, 2014, pp. 457–471. Available: http://dx.doi.org/10.1007/978-3-319-11964-9_29

[19] P. Ferragina and U. Scaiella, "Fast and Accurate Annotation of Short Texts with Wikipedia Pages," IEEE Softw., vol. 29, no. 1, pp. 70–75, Jan. 2012. Available: http://dx.doi.org/10.1109/MS.2011.122

[20] D. B. Nguyen, J. Hoffart, M. Theobald, and G. Weikum, "Aida-Light: High-Throughput Named-Entity Disambiguation," Linked Data on the Web at WWW2014, 2014.

[21] R. Blanco, P. Boldi, and A. Marino, "Entity-Linking via Graph-Distance Minimization," in Proceedings 3rd Workshop on GRAPH Inspection and Traversal Engineering, GRAPHITE 2014, Grenoble, France, 5th April 2014., 2014, pp. 30–43. Available: http://dx.doi.org/10.4204/EPTCS.159.4

[22] R. S. Sinha and R. Mihalcea, "Unsupervised Graph-Basedword Sense Disambiguation Using Measures of Word Semantic Similarity." ICSC, vol. 7, pp. 363–369, 2007. Available: http://dx.doi.org/10.1109/icsc.2007.87

[23] B. Hachey, W. Radford, and J. R. Curran, "Graph-Based Named Entity Linking with Wikipedia," Web Information System Engineering–WISE 2011, pp. 213–226, 2011. Available: http://dx.doi.org/10.1007/978-3-642-24434-6_16

[24] Y. Rochat, "Character Networks and Centrality," Ph.D. dissertation, University of Lausanne, 2014.

[25] D. A. Smith and G. Crane, "Disambiguating Geographic Names in a Historical Digital Library," Research and Advanced Technology for Digital Libraries, pp. 127–136, 2001. Available: http://dx.doi.org/10.1007/3-540-44796-2_12

[26] S. Fernando and M. Stevenson, "Adapting Wikification to Cultural Heritage," in Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics, 2012, pp. 101–106.

[27] E. Agirre, N. Aletras, P. Clough, S. Fernando, P. Goodale, M. Hall, A. Soroa, and M. Stevenson, "Paths: A System for Accessing Cultural Heritage Collections." in ACL (Conference System Demonstrations), 2013, pp. 151–156.

[28] M. De Wilde, "Improving Retrieval of Historical Content with Entity Linking," New Trends in Databases and Information Systems, pp. 498–504, 2015. Available: http://dx.doi.org/10.1007/978-3-319-23201-0_50

[29] F. Moretti, Graphs, Maps, Trees: Abstract Models for a Literary History. Verso, 2005. Available: https://books.google.fr/books?id=YL2kvMIF8hEC

[30] M. Jockers, Macroanalysis: Digital Methods and Literary History, ser. Topics in the Digital Humanities. University of Illinois Press, 2013. Available: https://books.google.fr/books?id=mPOdxQgpOSUC

[31] F. Ciotti, M. Lana, and F. Tomasi, "TEI, Ontologies, Linked Open Data: Geolat and Beyond." Available: https://jtei.revues.org/1365

[32] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," in Proceedings of the 16th International Conference on World Wide Web, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 697–706. Available: http://dx.doi.org/10.1145/1242572.1242667

[33] W. A. Gale, K. W. Church, and D. Yarowsky, "One Sense per Discourse," in Proceedings of the Workshop on Speech and Natural Language, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 233–237. Available: http://dx.doi.org/10.3115/1075527.1075579

[34] C. Brando, F. Frontini, and J.-G. Ganascia, “Linked Data for Toponym Linking in French Literary Texts,” in Proceedings of the 9th Workshop on Geographic Information Retrieval, ser. GIR ’15. New York, NY, USA: ACM, 2015, pp. 3:1–3:2. Available: http://dx.doi.org/10.1145/2837689.2837699

[35] V. Bryl, C. Bizer, R. Isele, M. Verlic, S. G. Hong, S. Jang, M. Y. Yi, and K.-S. Choi, “Interlinking and Knowledge Fusion,” Linked Open Data – Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project, pp. 70–89, 2014. Available: http://dx.doi.org/10.1007/978-3-319-09846-3_4

[36] J.-F. Baget, M. Chein, M. Croitoru, J. Fortin, D. Genest, A. Gutierrez, M. Leclère, M.-L. Mugnier, and E. Salvat, “RDF to Conceptual Graphs Translations,” in 3rd Conceptual Structures Tool Interoperability Workshop: 17h International Conference on Conceptual Structures, ser. LNAI, no. 5662. Moscow, Russia: Springer, Aug. 2009, p. 17.

[37] L. C. Freeman, “A Set of Measures of Centrality Based on Betweenness,” Sociometry, pp. 35–41, 1977. Available: http://dx.doi.org/10.2307/3033543

[38] F. Frontini, C. Brando, and J.-G. Ganascia, “REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions,” in Digital Humanities 2016: Conference Abstracts, ADHO. Jagiellonian University & Pedagogical University, 2016, pp. 193–197. Available: http://dh2016.adho.org/abstracts/362