HAL
open science

Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts

Francesca Frontini, Carmen Brando, Jean-Gabriel Ganascia

To cite this version:

Francesca Frontini, Carmen Brando, Jean-Gabriel Ganascia. Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts. First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, Arnaud Zucker; Isabelle Draelants; Catherine Faron Zucker; Alexandre Monnin, Jun 2015, Portorož, Slovenia. hal-01203358

HAL Id: hal-01203358

https://hal.science/hal-01203358v1

Submitted on 22 Sep 2015

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.

Semantic Web Based Named Entity Linking for Digital Humanities and Heritage Texts

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

[1] Labex OBVIL. LiP6. CNRS, 4 place Jussieu, 75005, Paris,
{Francesca.Frontini,Carmen.Brando,Jean-Gabriel.Ganascia}@lip6.fr
[2] Istituto di Linguistica Computazionale CNR, Pisa, Italy,
{Francesca.Frontini}@ilc.cnr.it

Abstract. This paper proposes a graph based methodology for automatically disambiguating authors' mentions in a corpus of French literary criticism. Candidate referents are identified and evaluated using a graph based named entity linking algorithm, which exploits a knowledge-base built out of two different resources (DBpedia and the BnF linked data). The algorithm expands previous ones applied for word sense disambiguation and entity linking, with good results. Its novelty resides in the fact that it successfully combines a generic knowledge base such as DBpedia with a domain specific one, thus enabling the efficient annotation of minor authors. This will help specialists to follow mentions of the same author in different works of literary criticism, and thus to investigate their literary appreciation over time.

Keywords: named-entity linking, linked data, digital humanities

1 Introduction

Named Entities (NE) are linguistic expressions that stand like rigid designators for referents; such entities normally include names of persons, geographical places, organizations, but also temporal references such as dates. Enriching mentions with a link to its referent by means of a unique identifier is crucial for the semantic annotation of texts. This is done by pointing to an external resource, such as a Universal Resource Identifier (URI) in the Linked Open Data (LOD) cloud. Segments in text referring to a Named Entity are known as entity mentions.

Named Entity Linking (NEL) [9] is a sub task of Named Entity Recognition and Disambiguation (NERD). NERD algorithms automatically detect entities in texts and assign them to a given class[3]. The NEL module assigns a unique identifier to the detected entities, thus disambiguating them by pointing to their referent. Linking is crucial since the same mention can represent different entities in different contexts and at the same time one entity can be mentioned in the text in different forms. So for instance the mention "Goncourt" can refer

[3] See [8] for a survey on NER.

to any of the two Goncourt brothers, Edmond or Jules. At the same time Jules de Goncourt can be referred to in the text as "Goncourt", "J. Goncourt", "J. de Goncourt", ... This means that, 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 a Named Entity 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. The bibliographic identifier "Goncourt, Jules de (1830-1870)", as well as the links <http://www.idref.fr/027835995> and <http://fr.dbpedia.org/page/Jules_de_Goncourt> are examples of such an identifier.

Besides ensuring disambiguation, linking also performs an important additional task, namely textual enrichment, in that it connects the mention with sources of additional information - such as DBpedia in the previous example - that needs not be stored in the text but can be accessed when required. In the case of Edmond de Goncourt, additional information from DBpedia can tell us what books he authored, where he was born, ...

The main issue with NEL in digital humanities is that mentions of persons often refer to individuals that are not listed in general ontologies such as Yago or DBpedia, that constitute the typical knowledge base for linking in other domains. Such individuals are often present in other knowledge bases, notably bibliographical linked data repositories (such as the French National Library BnF linked data repository). On the other hand, linking requires access to ontological knowledge, in that choosing between two individuals having the same name may requires comparing the context of the mention with a priori knowledge. In this respect, knowledge bases such as DBpedia remain an important source of general knowledge of the World. Thus the ideal linking algorithm for literary criticism texts combines general and domain specific sources. The experiment here described goes in this direction.

The paper will first present previous approaches to NEL, then the proposed graph based disambiguation algorithm based on the notion of centrality, finally describe the experiment carried out on the corpus and the results. Some conclusions and suggestions for further improvement of the algorithm are finally given.

2 Previous approaches

Previous approaches for NEL can be divided in two main families. Those using text similarity and those using graph based methods. 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 [7], that 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 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.

Graph-based approaches rely on formalised knowledge described in graph form that is built from a Knowledge Base (KB) (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. A reader may decide that the mention "James" refers to philosopher "William James" and not to writer "Henry James" because it occurs in the same context as "Hume" and "Kant". In the same way such algorithms build a graph out of the candidates available for each possible referent in a given context and 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 a paragraph) containing possibly more than one mention, so that the disambiguation of one mention is helped by the other ones.

This kind of approach is similar to the one used in Word Sense Disambiguation [11], where a set of words in a given sentence needs to be labeled with the appropriate sense label by using the information contained in a lexical database such as WordNet. The key idea of this approach is that for all ambiguous words in the context, senses that belong to the same semantic space should be selected, and that in this way two ambiguous words can mutually disambiguate each other. More specifically, a subgraph is built, constituted only of the relevant links between the possible senses of the different words, and then for each alternative sense labeling, the most central is chosen. This procedure, when applied to such context specific subgraphs, ensures that in the end the chosen senses for each word will be the one better connected to each other.

Centrality is an abstract concept, and it can be calculated by using different algorithms[4]. In [11] 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 similarity measures. While the combination of all centrality algorithms scores the best, Indegree centrality seems to be the better performing when compared to the other ones in terms of precision.

This graph based approach has been applied to NEL, where mentions take the place of words and Wikipedia articles that of WordNet synsets. Here too centrality measures are performed on the Wikipedia structure in order to use the rich set of relations to disambiguate mentions. More specifically in [4] English texts were disambiguated using a graph that relies only on English Wikipedia, and was constituted of the links and of the categories found in Wikipedia articles. So for instance the edges of the graph represent whether ArticleA links

[4] For a discussion of the notion of centrality see also [10]

to ArticleB or whether ArticleA has CategoryC. Here too "local" centrality is then used to assign the correct link to the ambiguous mention. We have chosen a graph-based approach to NEL that will be described in the next section.

3 Our approach

Our approach to disambiguate NE mentions is a graph-based one. Vertices are represented by URIs of mention candidates (e.g. dbpedia:Victor_Hugo) as well as URIs of concepts (e.g. foaf:Person) or individuals connected to at least two different candidates. Edges are semantic relations defined explicitly between URIs (e.g. "type"). The graph is undirected and their vertices and edges are *a priori* unweighted. We take advantage of the notion of centrality in Graph Theory to link a NE mention with the URI of the most probable candidate for that mention. In other words, we want to find the subset of vertices of different candidates having the greatest number of edges among them. The edges and vertices of the graph are built leveraging knowledge from different LOD sources whose nature is graph-based.

We illustrate the proposed approach with an example. Let us consider the following phrase of a French text of literary criticism written by Albert Thibaudet (1936) :

Quant au rythme, si ***Victor Hugo*** *a dépassé* ***Lamartine****, il n'a pas été plus loin que* ***Vigny****.*

In bold there are three mentions automatically recognized by a NER algorithm, that need now be linked to an identifier.

For each mention, the NEL algorithm selects possible candidates by exact string matching of the current mention and dictionary entries (e.g. Hugo, M. Hugo) and retrieves the corresponding URIs of the listed LOD sources. An excerpt of the candidates of the three named-entities from the example is listed below by distinguishable personal information instead of URI for readability sake.

Candidates (Victor Hugo) = Hugo, Victor (1802-1885)

Candidates (Lamartine) = Lamartine, Alix de (1766-1829), Lamartine, Alphonse de (1790-1869), Lamartine, Elisa de (1790-1863)

Candidates (Vigny) = Vigny, Joseph Pierre de (1742-1812), Vigny, Benno (1889-1965), Vigny, Alfred de (1797-1863)

Thanks to the URIs, it is possible to retrieve from the Web of Data the associated RDF graph for each candidate and combine them into a single graph. It should contain only those predicates involving at least two candidates of different mentions because we only want the predicates that play an important role in the disambiguation process. Calculating the centrality for every candidate will then give us the best candidates for the three mentions. Figure 1 shows an excerpt of the resulting graph where the chosen mention candidates are marked in bold. We can notice that the vertex yago:RomanticPoets is the one that influences the centrality measure the most because it is shared by the three chosen

candidates. Likewise, other vertices connected to the chosen nodes, such as dbpedia:romanticisme and dbpedia:Alexandru_Macedonski, are influential.

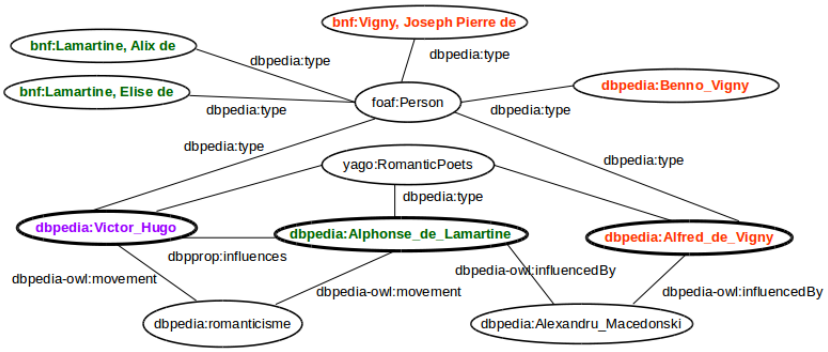

Fig. 1. Excerpt of the chosen URIs (in bold) for three candidates; a color designates all candidates of a single mention.

Named Entities are disambiguated and referenced within the context of a paragraph, so in principle two (identical) mentions of the same author within one paragraph will always receive the same link, while the same mention in different paragraphs might be assigned a different referent, depending on the other mentions it occurs with.

The NEL task is commonly defined in such a way that it does not assume the existence of the correct referent among the candidates in the knowledge base [5]. This is due to the fact that Wikipedia/DBpedia can hardly be a complete knowledge base even for textual genres such as contemporary newspapers articles. This seems even less true for the corpus that constitutes the object of our experiment. French literary criticism texts contain references not only to famous authors, but also to other minor figures that are not listed in Wikipedia. Therefore our proposal is to aim for a quasi complete reference base for the task of referencing authors.

Our approach relies importantly on a lookup dictionary; this is the subject of the following section.

4 LOD-based lookup dictionary

Linked data [1] is an important way of publishing knowledge in the Semantic Web. Such data is easily available via web services; LOD is composed of triplets of the form (subject, predicate, object) where subjects designate URIs, objects may be URIs or data-typed literals, and predicates represents binary relations. Queries can be run in the SPARQL language and data is provided with a dereferenciable and persistent identifier called URI (Uniform Resource Identifier). Many

of the available linked data are of great interest for digital humanities [12], and for the domain of literary criticism in particular. More specifically, information on authors for French texts can be found in the French version of DBpedia on the one hand, and in the catalogue of the Bibliothque Nationale de France (BnF) on the other.

The French DBpedia is constituted of the articles of the French version of Wikipedia. In DBpedia entries are classified one or more of the types of the DBpedia ontology. So for instance the author known as Stendhal[5] is classified as *Person*, *Artist*, *Writer*, and at the top level, as *Thing*. Moreover, authors are linked to each other by horizontal relations such as *InfluencedBy*, and, indirectly, by being linked to the same concept, such as *Romanticism*. BnF entries list all authors of books ever published in France; their entries contain information on date of birth and death, gender, alternative names, works authored. For instance the BnF entry for Voltaire[6] gives several alternative names such as Franois-Marie Arouet (Voltaire's real name), Wolter, Good Naturd Wellwisher, ...

Most crucially, BnF links its entry to the DBpedia one when existing, thus making it very easy to connect the two resources in one knowledge graph. Moreover, BnF entries also list the author's Idref, which is the official identification system used by French universities and higher education establishments to identify, track and manage the documents in their possession. The combination of these two sources was considered able to grant a sufficient coverage for a corpus of French literary criticism, thus the BnF and the DBpedia SPARQL endpoints were queried for all authors, retrieving their biographic information (name, surname, alternative names, dates of birth and death, title, ...) in structured form.

In order to be able to retrieve all possible mentions of an author, this information was processed into a dictionary of authors, that contains all alternative names of an author, plus a series of alternative forms automatically generated, with the links to BnF and DBpedia entries. Automatically generated alternative names are of the form:

- surname only (Rousseau)
- initials + surname (J.J. Rousseau, JJ Rousseau, ...)
- title + surname (M. Rousseau, M Rousseau)

Gīvėň the domain (French literature) this procedure ensures that 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.

5 Implementation of the NEL algorithm

The NEL algorithm processes a file in XML-TEI format[7]; NE mentions are annotated with NER annotations (e.g. tag <persName>) for every paragraph; the

[5] http://fr.dbpedia.org/page/Stendhal
[6] http://data.bnf.fr/11928669/voltaire/
[7] http://www.tei-c.org/index.xml

algorithm is devised to processes one single class at a time (here Person). It uses a lookup dictionary per class listing superficial forms and their associated URIs from LOD sources, as described in the previous section. The algorithm produces an enriched version of the input file indicating the chosen candidate for each mention. We developed our implementation in Java ; RDF data is processed thanks to the Jena API[8]; graphs are manipulated by the JgraphT API[9] and implementation of centrality measures are available in the Social Network analysis tool, JgraphT-SNA[10]. In particular, the algorithm performs the following steps for every paragraph of the XML-TEI file:

1. look for URIs of mention candidates in the dictionary
2. retrieve the RDF graphs of those URIs
3. simplify and combine graphs then compute the selected centrality measure
4. choose URI of candidate with the higher score per mention then write results in TEI file

The algorithm searches for (1) possible candidates of mentions by exact string matching the mentions of the current paragraph and superficial forms in the dictionary; there must be at least one ambiguous mention to continue. It retrieves URIs (BnF, DBpedia) of mention candidates from dictionary entries. Next, the RDF graph is retrieved (2) for every URI and converted to a JgraphT-compatible graph, where RDF objects and subjects are vertices and RDF predicates are edges. Irrelevant edges and vertices are removed from graphs. We keep edges which involve at least two vertices representing URIs candidates. Information coming from different sources is combined into a single graph (3); the way we combine graphs is straightforward. The fusion is implicitly done thanks to one of the main LOD principles which consists of reusing vocabularies published in the LOD vocabulary cloud. In other words, edges (predicates) and vertices (URI nodes) should be shared by at least two graphs associated to candidates of different mentions. The selected centrality measure (e.g. closeness) is calculated for the resulting graph. Finally, the algorithm chooses (4) the URI of the mention candidate with the higher centrality score and annotates the input XML-TEI file with this information.

Furthermore, simplification of graphs and calculation of centrality measures in the combined graph are crucial parts of the algorithm (3). This step is detailed in the Algorithm 1. It essentially removes edges which are irrelevant to calculate a centrality measure, in other words, it deletes those edges which involve at most one vertex of a non-candidate URI.

6 Experiments and results

This section describes the experiments settings used to test our proposal as well as preliminary results which are encouraging. In this experiment, in order to

[8] https://jena.apache.org/
[9] http://jgrapht.org
[10] https://bitbucket.org/sorend/jgrapht-sna

Algorithm 1 NEL: simplify and combine graphs, compute centrality

```
Require: graphs: graphs of candidates per mention, measure: centrality measure
  for graph in graphs do
    initialize vertexToDelete
    for vertex in graph do
      if vertex is not a candidate then
        initialize vertexCheck
        for edges of vertex do
          if vertex1 notEqual vertex AND vertex1 is candidate then
            vertexCheck.add(vertex1)
          end if
          if vertex2 notEqual vertex AND vertex2 is candidate then
            vertexCheck.add(vertex2)
          end if
        end for
        if size of vertexCheck < 2 then
          vertexToDelete.add(vertex)
        end if
      end if
    end for
    graph.removeAllVertices(vertexToDelete)
    chosenURIs = calculateCentrality(measure, graph)
  end for
  return chosenURIs, chosen candidate per mention
```

evaluate the performance of the algorithm, the linking is performed on correctly identified and classified authors.

6.1 Experiment settings

The test corpus consists of a French text of literary criticism titled "Une thése sur le symbolisme" (A thesis about Symbolism) and it is the first volume of the work named "Réflexions sur la littrature" (Reflexions on literature) published by Albert Thibaudet in 1938.

The text is drawn from a larger "Corpus critique"[11], published in TEI by the Labex OBVIL and containing a large collection of critical essays by different authors.

The chosen text in particular presents a high density of authors' mentions, so that each paragraph generally contains an average of 2-3 mentions that are treated at the same time by the algorithm. Mentions concerning authors were manually annotated by two experts in French literature; URIs assigned to mentions are those from Idref[12]. Guidelines to manual annotation were those proposed by the MUC7 conferences as well as those defined by the XML/TEI standard. The resulting test corpus contains 1021 manually annotated mentions of

[11] http://obvil.paris-sorbonne.fr/corpus/critique/
[12] www.idref.fr

person entities. We measure the precision of the proposed NEL approach in terms of the attribution of the right URI to a mention with respect to the URI manually assigned by humans. The authors lookup dictionary was automatically built in advance thanks to the BnF LOD source which is rich in SameAs predicates pointing to DBpedia and Idref URIs. The resulting lookup dictionary is composed of 4,218,798 author names including their alternative names (e.g. M. Lamartine, Monsieur Lamartine, etc.). We chose 3 centrality measures commonly used in social network analysis and the word-sens disambiguation problem, these are: *DegreeCentrality*[3], *BrandesBetweennessCentrality*[2], *FreemanClosenessCentrality*[3], as implemented in the JgraphT-SNA tool.

6.2 Results and Analysis

The test results with the three algorithm are shown in table 1,

Table 1. Results with different centrality measures on test corpus.

Centrality Measure Used	Precision	Unassigned Links
DegreeCentrality	0.73	23
BrandesBetweennessCentrality	**0.74**	23
FreemanClosenessCentrality	0.43	23

Precision is calculated comparing the number of correctly assigned links over the total of manually annotated entities of authors. The best result is obtained with BrandesBetweennessCentrality, with a precision of 0.74. DegreeCentrality has a comparable performance, FreemanCloseness centrality seems to heavily underperform with respect to the other centrality measures. The last column of table 1 shows the number of empty links over the total.

These first results are satisfying: though far from the 85% accuracy that is normally achieved by similar algorithms on the news domain, such levels of precision are nevertheless remarkable, considering that in many cases the text discusses minor authors, today unknown, that are not necessarily listed in DBpedia. Moreover, the use of BnF makes the number of candidates (and thus the possibility of error) explode, with sometimes as much as 20 or more possible candidate for a mention.

To quantify authors incompleteness in both the DBpedia and BnF data sets used in this experiment, we count the number of mentions in which the algorithm (using DegreeCentrality measure) does not find any corresponding URI in the chosen KB. In this manner, there are 160 author mentions, out of 1021 mentions identified in the corpus by the algorithm, that have no match in DBpedia, that is around 16%. Remarkably, there are only 23 mentions (i.e. 2%) that have no match in either BnF or DBpedia. Notice that all authors in this test set that are in DBpedia are also in BnF.

The most frequent mistakes considering DegreeCentrality and BrandesBetweennessCentrality measures (the most similar and precise ones) concern the

following authors: Vielé-Griffin, Francis (1864-1937); Boileau, Nicolas (1636-1711); Barrès, Maurice (1862-1923); Payen, Fernand (1872-1946); Lefranc, Abel (1863-1952); Shakespeare, William (1564-1616); Spencer, Herbert (1820-1903); Goncourt, Edmond de (1822-1896) and brother Goncourt, Jules de (1830-1870); Mentré, Franois (1877-1950). The algorithm makes three types of mistakes.

MISSING CANDIDATES - In 23 cases the algorithm is unable to retrieve any candidate from the lookup dictionary, since the author is not present in any knowledge base. This is the case of author Francis Vielé-Griffin. In other cases the correct entity is present but not associated with the required pseudonym. This is the case of William Shakespeare's alleged alter ego William Stanley[13]. This alias is not listed in the dictionary for Shakespeare, therefore, it is not possible to assign both mentions to the same person (and thus the same URI).

MISSING CONTEXT - In some rare cases only one ambiguous author's mention is present in a single paragraph, thus the algorithm resorts to a fall back strategy, choosing the entity with more links in absolute. Sometimes this strategy causes errors, as in the case of "Vigny", for whom, in isolation, the wrong link to Auriane Vigny is chosen.

INCOMPLETE INFORMATION - In some cases the context of the sentence should be sufficient to produce a correct disambiguation but the NEL algorithm makes mistakes due to lack of links in the knowledge base, which prevents the centrality measure to produce the desired result. For instance, "Shakespeare", when mentioned in the context of Shakespearian critic Abel Lefranc, should produce the correct linking to William, but Nicolas is chosen instead. Clearly explicit links between Abel Lefranc and the object of his studies are missing in the knowledge bases. Ancient authors also tend to cause problems due to lack of information, e.g. the Greek author Lysias is mistaken for an homonymous French revolutionary collective.

WRONG, MISLEADING INFORMATION - Sometimes the knowledge bases contain wrong or misleading information. For instance there exist a BnF entry for the "Ronsard family", classified as foaf:Person, which is chosen instead the correct assignment, namely one of its members, Pierre de Ronsard. The opposite is also true, so some mentions refer to both Goncourt brothers as a collective noun, but the algorithm chooses one of the two. Finally, wrong or misleading pseudonyms are sometimes listed in BnF for an author, causing wrong candidates to be injected in the graph and sometimes selected. So for instance "Descartes" is listed as a pseudonym for novelist Horace Walpole and thus sometimes Walpole is wrongly chosen as the link for philosopher Descartes.

Error analysis also shows that sometimes relevant information that is present in the knowledge base is not used in the decision process because it cannot be encoded in the graph in the form of links. A typical example is temporal information which is encoded in the form of dates (data-typed literals). In other words the fact that - for a given context - two candidate referents lived in the same period of time cannot be taken into account.

[13] Stanley is believed by some to be the real author behind Shakespeare's works.

To evaluate the impact of the temporal dimension, we chose to evaluate against an index from which we removed authors born after the date of publishing of the work. The results show a slight improvement with **DegreeCentrality reaching precision 0.78** and BrandesBetweennessCentrality 0.77. A greater improvement may be obtained using a more sophisticated graph building algorithm, that transforms information about dates of birth and death in links that can connect authors in a measurable way.

7 Conclusion and future work

We presented an algorithm to perform NEL on a corpus of 19th century literary criticism, with the specific goal of disambiguating and referencing author mentions for research purposes. The NEL module is meant to be used in combination with a NER module, and will help researchers in the creation of digital literary editions enriched with information about authors. The main purpose of this work is to help scholars in history of literature to perform complex queries in order to study the literary appreciation of authors over time, and investigate the history of literary criticism in French literature. More specifically the enrichment of the aforementioned "Corpus critique" is meant to enhance ongoing research in the history of scientific ideas, and to provide a way to follow the dissemination of theories and concepts defined by Charles Darwin, Claude Bernard, Henri Bergson in non scientific texts of their time.

The reported experiment shows how combining different sources can be useful to perform linking on a domain specific corpus with satisfying results. While the precision is not yet state of the art, it is nevertheless remarkable, considering that it is the first time graph that centrality algorithms have been used for NEL combining DBpedia with a domain specific source. Tests showed significant differences between one implementation of centrality and the other two. Error analysis suggests possible improvements of the algorithm, including the ad hoc transformation of temporal information - present in the knowledge base in the form of literals - into links of the context graph. Another possible evolution of the algorithm would be to assign different weights to the edges so that for instance sharing the same literary circle becomes a more important relation than being born in the same town. Weights would be learned from manually annotated data. Further experiments will be carried out with different corpora and on different categories of entities, notably places.

Experimenting with the size of the context will also be necessary, in order to find the best trade-off between efficiency and informativeness. A more ample context (ideally a whole chapter) may produce a better graph of candidates, such that all mentions can disambiguate each other correctly. But at the same time this may introduce noise, and also generate a graph so big that its construction and the calculation of centrality may require too much time.

Another possible evolution of the algorithm could be to improve the graph fusion procedure. So far, our strategy does not handle the proper fusion of individuals which are described heterogeneously by the different sources (e.g. Victor

Hugo as described by BnF, as described by DBpedia, and so on). In this study we chose to study the problem from a quantitative point-of-view and thus to consider existent knowledge as it is without a pre-processing step. In the future, we foresee to make use of strategies commonly applied in Conceptual Graphs for information fusion [6]. In this way, the resulting graph would better concentrate domain knowledge (i.e. avoid redundancy and conflicts) and thus calculate a more accurate centrality measure.

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. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. 5(3), 122 (2009)
2. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25(2), 163–177 (2001)
3. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
4. Hachey, B., Radford, W., Curran, J.R.: Graph-based named entity linking with wikipedia. In: Web Information System Engineering–WISE 2011, pp. 213–226. Springer (2011)
5. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R.: Evaluating entity linking with wikipedia. Artificial intelligence 194, 130–150 (2013)
6. Laudy, C., Ganascia, J.G.: Information fusion using conceptual graphs: a tv programs case study. In: ICCS. pp. 158–165 (2008)
7. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. ACM (2011)
8. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
9. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Springer (2013)
10. Rochat, Y.: Character Networks and Centrality. Ph.D. thesis, University of Lausanne (2014)
11. Sinha, R.S., Mihalcea, R.: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: ICSC. vol. 7, pp. 363–369 (2007)
12. Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. Literary and linguistic computing (2013)