Domain-adapted named-entity linker using Linked Data

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

[1] Labex OBVIL. LiP6, UPMC. 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. We present REDEN, a tool for graph-based Named Entity Linking that allows for the disambiguation of entities using domain-specific Linked Data sources and different configurations (e.g. context size). It takes TEI-annotated texts as input and outputs them enriched with external references (URIs). The possibility of customizing indexes built from various knowledge sources by defining temporal and spatial extents makes REDEN particularly suited to handle domain-specific corpora such as enriched digital editions in the Digital Humanities.

Keywords: named-entity disambiguation, evaluation, linked data, digital humanities

1 Introduction

Semantic annotation plays an important role in the construction of enriched digital editions in the domain of Digital Humanities. The TEI annotation standard[3] defines specific tags to mark-up key words in texts and to link them to background knowledge by using external references. Thus, for instance, it is possible to annotate topics, citations, places, organization and person names. References can point to internal identifiers or to external resources that are available in the form of Linked Data (LD) on the semantic Web. Such annotations can be used to facilitate research on the enriched texts (by allowing the creation of indexes and the performing of advanced queries). Moreover, they can also enhance the utilization of texts by the final user, since TEI can be easily converted into other formats, notably Electronic Publication (EPUB). Physical supports such as ebook readers could in future make use of the enriched information in the form of external links.

The results of the work described here make up part of an editorial pipeline the aim of which is produce an enriched digital edition of a corpus of French literary criticism and essays from the 19th century[4]. Such a corpus, once completed, is intended to allow researchers to gain a more comprehensive perspective on

[3] http://www.tei-c.org/index.xml
[4] http://obvil.paris-sorbonne.fr/corpus/critique/

the evolution of ideas both in literature as well as in the sciences over the years. While the production of high quality digital editions requires manual checking, natural language processing techniques can speed up the process to a great extent. In particular, the detection of person mentions is a very important step for the enrichment of such a corpus, as it allows researchers to trace references to authorities (authors, scientists, critics, etc.) over time and from different books.

While existing Named Entity Recognition and Classification (NERC) algorithms can help human beings to detect persons' mentions, previous experience has shown that this step is not particularly difficult for trained annotators. Conversely, disambiguation and linking of mentions to an identifier can often be painstaking when done manually, as it requires searching external sources, such as mainstream online encyclopedias (Wikipedia) but also domain-specific databases, such as library databases. Named Entity Linking (NEL) algorithms[6] can thus be of great use, as they automatically assign an external identifier to already detected mentions, performing disambiguation and semantic annotation at the same time. They rely on external sources of structured or unstructured information to perform this task which is complex, because an entity is usually mentioned in the text in ambiguous forms.

While current NEL algorithms perform well on the contemporary news texts domain adaptation is, as always in NLP, an issue. Nineteenth century essays for instance often mention persons that were well known at the time but are currently hard to identify even for specialists. Thus although existing NEL tools are very efficient they mostly rely on generic background knowledge such as Wikipedia or DBpedia that may not be sufficient to grant ample coverage for all domains.

The solution we present here is a NEL tool, REDEN, that makes use of state of the art graph-based algorithms and at the same time can easily leverage knowledge from different sources (both generic and domain specific). REDEN is applied to already detected mentions, and relies natively on knowledge drawn from linked data sources for disambiguation, requiring only an index of superficial forms and URIs to be used as candidates. Such indexes can be built out of different sources using crawlers.

The purpose of this paper is to show that such a solution can be efficiently used for texts in which a domain specific knowledge base is necessary. We evaluate REDEN on ad-hoc test-sets from our corpus of essays and compare the results with those obtained with a freely available NEL tool. We first present the current state of the art in NEL, then briefly describe the tool and then focus on the index building facilities; some experiments are presented to compare our proposed solution to the state of the art, and we conclude with some remarks and ideas for future developments.

2 Related work

Supervised methods usually perform very well in NLP tasks but they rely on pre-annotated corpora for training which are not available for specific domains

such as French Literature and the Digital Humanities more generally. Therefore we concentrate on existing non-supervised approaches for NEL. These can be broadly divided into two main families. Those using text similarity along with statistical models and those using graph-based methods.

The best known tool from the first group is DBpedia Spotlight (DBSL) [2] which performs NER and DBpedia linking at the same time. It is based on cosine similarities and a modification of TF-IDF weights. More specifically, recognition of possible annotations in a text is performed using indexes built from the Wikipedia dumps and DBpedia by substring matching algorithms. DBSL generates DBpedia candidates for each possible annotation in two possible ways, using a language-dependent implementation (available for English and Dutch and more NLP-adapted) or a language-independent implementation which is more efficient but less precise. DBSL then selects the best candidates from the set of annotations by computing scores that combine statistical methods, the Wikipedia article links along with their anchor texts and textual context. This or similar methods are known to be very efficient, but they can are strongly dependent on the availability of textual descriptions for each entry.

Graph-based approaches rely on formalised knowledge in the form of a graph built from a knowledge base (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 available candidates 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 (e.g. paragraph) containing possibly more than one mention, so that the disambiguation of one mention is helped by that of the others. Hybrid approaches such as [9] can use both graph-based algorithms, more specifically graph-search ones, and text similarity measures; they rely on a single LD source.

The graph-based NEL is similar to graph-based Word Sense Disambiguation (WSD)[8], 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 field 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 those that are better connected to each other.

Centrality is an abstract concept, and it can be calculated by using different algorithms[5]. In [8] the experiment was carried out using the following algorithms:

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

Indegree, *Betweenness*, *Closeness*, *PageRank*, as well as with a combination of all these metrics using a voting system. Results showed the advantages of using centrality with respect to other similarity measures.

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 large set of relations to disambiguate mentions. More specifically in [5] English texts were disambiguated using a graph that relied only on English Wikipedia, and was made up of the links and of the categories found in Wikipedia articles. So for instance the edges of the graph represented whether ArticleA links to ArticleB or whether ArticleA has CategoryC. Here too "local" centrality is used to assign a correct link to an ambiguous mention. Such a WSD based tool for NEL is proposed by the DBpedia Graph Project[6]. However, it is highly dependent on DBpedia structure and only links to this broad-coverage data set and not to other domain-specific ones.

3 Our disambiguation approaches

With REDEN, we propose a graph-based, centrality-based approach that is particularly adapted for digital humanities; in fact it can exploit RDF sources directly (thus allowing users to choose domain adapted knowledge bases) and takes TEI-annotated texts as input. The disambiguation algorithm processes a file where NE mentions are already annotated (e.g. using the tag <persName>). Possible referents from all mentions in a given context and for a given class (e.g. person) are retrieved from an index built on the fly from selected LD sources. The fusion of homologous individuals across different sources is performed thanks to *owl:sameAs* or *skos:exactMatch* predicates; a graph is created where RDF objects and subjects are vertices and RDF predicates represent edges. Irrelevant edges are removed from the graph before calculating centrality: only edges which involve at least two vertices representing candidate URIs are preserved. Thanks to the selected centrality measure, the best connected candidates for each mention are chosen as referents and an enriched version of the input TEI file is produced.

To illustrate how REDEN works, we apply our approach to a paragraph from a French text of literary criticism entitled "Une thèse sur le symbolisme" (1936) written by Albert Thibaudet (1874-1936)[7]. Figure 1 shows an excerpt of the resulting graph where the chosen and correct mention candidates out of the nine present in the text are marked in bold. We can observe that the vertices *yago:Symbolist Poets*, *dbpedia:Charles_Baudelaire*, *dbpedia:French Dramatists And Playwrights* are the ones influencing the centrality measure the most.

[6] https://github.com/bernhardschaefer/dbpedia-graph but related published work is not available at this time

[7] Full text can be found here: http://obvil.paris-sorbonne.fr/corpus/critique/thibaudet_reflexions.html

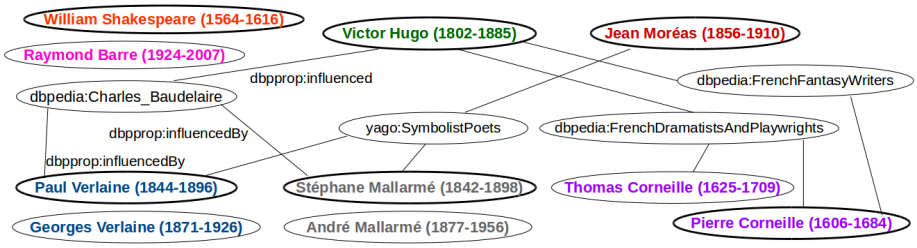

Fig. 1. Excerpt of the chosen URIs (in bold) for the candidates of the nine mentions; a color distinguishes candidates of a single mention; all edges represent *rdf:type* links except those marked differently.

We developed our implementation in Java[8]. In [4, 1], we described the steps of the algorithm in detail. In the following section, we detail the software component for building on-the-fly a domain-adapted index of possible mention candidates.

4 Domain-adapted crawler of Linked Data

As previously described, REDEN identifies and looks for mention candidates' URIs in indexes built on the fly per class. So for instance, the entries of a person in the index correspond to superficial forms of that person's name and the corresponding URIs from different Linked Data sets. REDEN then retrieves the RDF data from the corresponding URIs found in the index for each selected source.

Clearly, a source can be more useful for certain domains than another, for instance the French National Library (BnF) linked data repository is the most appropriate source for a domain such as French literary criticism. BnF provides the catalog of authors for all books ever published in France; their entries contain information on dates of birth and death, gender, alternative names, works authored, etc. Most importantly, BnF provides equivalence links (*owl:sameAs*) to other sources such as DBpedia or IdRef. Thus, BnF can be chosen as the entry point for building an index of authors. In particular information on gender and full names can be used to generate further alternative names - for instance: surname only (Rousseau), initials + surname (J.J. Rousseau, JJ Rousseau, ...), title + surname (M. Rousseau, M Rousseau), ... Equivalence links may also be stored, so that later on REDEN can build fused graphs from different sources.

Likewise, using DBpedia as entry point an index of places can be assembled. For instance, the DBpedia entity Gare de Paris Saint-Lazare has the following name forms (familiar, official and old ones) - for instance: Gare Saint-Lazare, Gare Saint-Lazare (Paris), Embarcadère de l'Ouest, ...

[8] Code source and useful resources can be found here: https://github.com/cvbrandoe/REDEN

We have therefore developed an index-building module which allows for the configuration and the implementation of a crawler for a chosen LD source. The main purpose of this module is to facilitate the construction and the update of entity indexes which should be suitable for the domain of the exploited corpora. In particular, the temporal dimension and possibly the spatial dimensions can be used as filters. For instance when working with 19th century texts, a time filter can a priori exclude the currently well-known singer Thierry Amiel as a candidate for the mention 'Amiel' and thus enable the identification of the correct referent, well-known writer Henri-Frédéric Amiel.

The proposed module takes into account domain parameters such as the temporal and spatial dimensions. For instance, a time period can be specified with respect to the date of birth of the persons described in the corresponding linked data set, or to any other property of the individual (date of death, periods of activity). Concerning the spatial extent of the index, one may be interested in including only persons who were born or have lived in Europe because the corpora probably concerns authors with these features. A bounding box can be defined by specifying a rectangle consisting of four Latitude and Longitude coordinates[9]. It is also straightforward to define new domain parameters to filter for themes, such as Symbolism or Psychology.

From a technical point-of-view, a crawler performs SPARQL queries using a selected end point. The created indexes can benefit from text search engines which enable the optimal querying of candidates by REDEN and therefore reduce query processing time.

Currently REDEN comes with two already implemented crawlers: one to collect an index of place candidates from DBpedia and one to crawl an index of author candidates from the LD repository of the BnF, containing alternative names for every entity. In the case of authors, as BnF entries point to DBpedia equivalents when available, these two sources can be easily fused using the aforementioned algorithm. In the case of places, there is no need to fuse other LD sources although one may be interested in including a place gazetteer available in linked data such as Pleaides[10] when working with Ancient texts.

5 Tool assessment

In previous experiments [4, 1], we have discussed and evaluated our graph-based method, which allows for state of the art accuracy in NEL for the texts in our domain. Here we assess REDEN as a tool, thus not only in terms of accuracy, but also of adaptability and usefulness. For this reason we compare the performance of our linker with an available state of the art tool such as DBpedia Spotlight (DBSL). As our texts are in French, the language independent version of DBSL is used. Thus the comparison is not fair, as this version performs both NERC and NEL at the same time, while we run REDEN on a TEI text with a pre-checked

[9] Some Web mapping tools easily provide this kind of information: http://www.mapdevelopers.com/geocode_bounding_box.php

[10] http://pleiades.stoa.org

set of classified mentions. Nevertheless, the comparison can help to give an idea of the differences in coverage that our method offers.

Since REDEN treats one class at a time, DBSL too is called with the chosen class as a parameter. Moreover, we use the most relaxed configuration possible, accepting any matching results with no filter on support or confidence, to maximize recall. DBSL is installed locally and launched using French DBpedia as a knowledge base. As for REDEN, the chosen centrality measure is *DegreeCentrality*[3].

The manually linked test corpus consists of a French text of literary criticism entitled "Une thèse sur le symbolisme" (A thesis about Symbolism) published by Albert Thibaudet in 1936, and a scientific essay entitled "L'évolution créatrice" (Creative Evolution) written by Henri Bergson and published in 1907. We can consider two measures of accuracy: A1, the proportion of correct links over the mentions for which a link was chosen, and A2, the proportion of correctly assigned links over the number of mentions overall. The second measure considers missing links as mistakes, and can be seen as a measure for coverage[11].

5.1 Experiment 1: Locations

Even though REDEN's strength lies in the possibility of merging resources, we will first compare the two algorithms on French DBpedia alone[12]. Since DBSL cannot process more than a paragraph at a time, REDEN too is configured to use paragraphs as contexts for its graphs. We choose the task of linking location mentions, which, in our texts are rather sparse but not particularly domain-specific. A priori we expect most referents for locations' mentions to be present in the knowledge base.

Corpus	Mentions	REDEN					DBSL				
		Found	Correct	None	A1	A2	Found	Correct	None	A1	A2
Thibaudet	382	257	257	132	1.0	0.67	198	195	184	0.98	0.51
Bergson	58	49	49	9	1.0	0.84	42	41	16	0.97	0.70

Table 1. Comparing the performance of REDEN with DBpedia spotlight on our two corpora for Locations. None=link not found.

As shown by the results, both linking algorithms are comparable. They both achieve excellent correctness rates when they assign links, but they miss a certain number of places. Missing places occur often due to the lack of alternative names, such as "Lacédémone" (another name for Sparta), or from entries that are misclassified in DBpedia, such as "Berlin" which is classified as a Concept and not

[11] These measures may be considered to roughly correspond to precision and recall, though such terms seem less adapted to describe this type of task.

[12] REDEN and DBSL use versions of 05/03/15 and of 17/07/2014, respectively. Unfortunately, it is technically not possible for REDEN to access a snapshot of DBpedia RDF resources from that particular date.

as a Location (in the French version of DBpedia). REDEN achieves a better coverage even when using the same knowledge base, possibly because it better exploits the available alternate names, or because DBSL does not provide an answer when it is unable to find a correspondence between the textual context of a mention and the description of a candidate.

As previously stated this evaluation is meant as a comparison of the algorithms, as REDEN works on pre-detected mentions while DBSL has to perform detection too. That said, in this experiment DBSL analyses only paragraphs containing locations and we force the type "location" in input; in most cases when DBSL is missing a link the name of the mention is written in capitalized words and coincides exactly with the title of a DBpedia entry; thus it is possible to conclude that, in such cases, it is not the detection but the linking that fails.

5.2 Experiment 2: Persons

The second experiment aims to show how our tool can achieve a much better performance on this corpus thanks to the possibility of combining two knowledge bases to target a specific domain. Here, the more difficult task of identifying authors' mentions is evaluated. DBSL is run using French DBpedia as usual, REDEN is run on an index of authors built from French DBpedia + the BnF linked data. Moreover, we exploit all the fine-tuning options offered by REDEN: in order to best exploit contextual information we increase the disambiguation context, using the chapter for Thibaudet, which is a text with high mentions' density, and the whole book for Bergson because it is poorer in mentions[13]. Moreover, we filter the index by preventing authors who weren't born or were too young at the time of publication of each work from becoming candidates.

Corpus	Mentions	REDEN					DBSL				
		Found	Correct	None	A1	A2	Found	Correct	None	A1	A2
Thibaudet	1027	1004	878	23	0.87	0.85	177	174	850	0.98	0.17
Bergson	277	274	222	3	0.81	0.68	8	8	269	1.0	0.03

Table 2. Comparing the performance of REDEN with targeted settings with DBpedia spotlight on our two corpora for Persons.

As shown in Table 2, DBSL achieves a high correctness rate for persons when it makes a choice, assigning almost perfectly each mention found on both corpora. On the other hand, the amount of entities found is much smaller. The reason is obviously that REDEN exploits additional information, granting coverage for authors who aren't present in DBpedia.

At the same time, we notice that DBSL correctly recognizes the mention of (Alphonse de) "Lamartine" but not those of (Maurice) "Barrès" and (Alphred

[13] In [1], we showed that considering different context sizes can increasingly improve correctness rates; we also assess the impact of filtering the knowledge base with temporal window.

de) “Vigny”, when all three authors have a referent in DBpedia, and despite the fact that we are running the algorithm with the lowest possible confidence. This may mean that the unstructured information on which DBSL relies is not enough to make a decision, while the graph of relations that REDEN exploits is richer in information. Generally speaking REDEN’s correctness rate is close to state of the art (even with respect to other graph-based algorithms, whose stated correctness rate is around 0.85), although Thibaudet, which is richer in mentions, gives better results.

It is important to note here that DBSL is more precise than REDEN in those cases where it actually chooses a link (A1). In particular, it beats our method when a paragraph contains only one ambiguous candidate mention. In this case, the unstructured textual context is a useful source of information to disambiguate, while the graph-based algorithm cannot use background information. Nevertheless, the greater coverage of the graph-based approach makes the overall post-processing work of manual checking less cumbersome in terms of both correction and integration.

6 Conclusions and future work

In this article we presented REDEN, a graph-based tool for Named Entity Linking that can be easily extended to different online sources and tailored to suit various different needs. Domain adaptation is enabled by entity indexes built on the fly using linked data crawlers. New sources for a particular domain can be added by configuring a new crawler with the appropriate SPARQL query. Experiments involving the disambiguation of persons’ and locations’ names in a corpus of French essays of the 19th century show how the graph-based algorithm gives us state of the art correctness results while allowing for more flexibility and coverage.

The experiment with DBSL highlighted some of the practical advantages of our tool. First, DBSL is very dependent on the DBpedia. Configuring it to use new knowledge bases such as BnF does not yet seem possible. Furthermore, REDEN works natively with RDF. The structured information used has thus the potentiality to grow as new links are added to the semantic web. REDEN uses online sources (end points) which allows for the easily downloading of the most up-to-date version of the data from the corresponding repository. These data are cached by REDEN in order to download an RDF resource only once. On the contrary, DBSL uses offline resources which can be an advantage when bandwidth is limited, however updating the DBSL knowledge base looks to be a very difficult and time consuming process. Another technical limitation is the size of the context that DBSL uses for its decision; the tool runs as a Web service accepting HTTP calls and the URL size is restricted. In REDEN, the size of the context does not need to be fixed, but can be set by the user according to predefined textual partitions in TEI; thus the tool can run using all mentions in a paragraph, a chapter, or a whole text as a disambiguation context. Finally

crawlers can be customized in such a way as to apply (time or space) filters on candidates, thus making it domain adapted.

For these reasons we conclude that the proposed methodology seems better adapted to the task of linking named-entities to a knowledge base other than DBpedia. Thus it seems particularly useful for digital editions in the humanities, where texts are already in TEI format, and specialists have a good knowledge of the text and of relevant sources of information.

As prospective work we intend to perform an exhaustive comparison of RE-DEN to other graph-based NEL tools such as AGDISTIS[9] and the DBpedia Graph Project focusing on domain-specific test corpora, in particular, on texts used in French Literature and more generally in the Digital Humanities.

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. Brando, C., Frontini, F., Ganascia, J.G.: Disambiguation of named entities in cultural heritage texts using linked data sets (accepted). In: Proceedings of the First International Workshop on Semantic Web for Cultural Heritage in Conjunction with 19th East-European Conference on Advances in Databases and Information Systems (2015)
2. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
3. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
4. Frontini, F., Brando, C., Ganascia, J.G.: 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. pp. 77–88 (2015), http://ceur-ws.org/Vol-1364/
5. 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)
6. 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)
7. Rochat, Y.: Character Networks and Centrality. Ph.D. thesis, University of Lausanne (2014)
8. Sinha, R.S., Mihalcea, R.: Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: ICSC. vol. 7, pp. 363–369 (2007)
9. Usbeck, R., Ngonga Ngomo, A.C., Auer, S., Gerber, D., Both, A.: Agdistis - graph-based disambiguation of named entities using linked data. In: 13th International Semantic Web Conference (2014)