Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 137 (2018) 199–210

Procedia Computer Science

www.elsevier.com/locate/procedia

SEMANTiCS 2018 – 14th International Conference on Semantic Systems

Constructing a knowledge base for entity linking on Irish cultural heritage collections

Gary Munnelly, Séamus Lawless

Adapt Centre, Trinity College, Dublin

Abstract

A known problem for proponents of Entity Linking in Cultural Heritage collections is the poor representation of entities in popular knowledge bases such as DBpedia. This is largely due to the niche nature of the entities contained in these collections. Where such problems arise it may be possible to generate more complete knowledge bases from existing resources used by scholars who specialise in the domain of the collection in question. This paper presents the process by which a knowledge base for informing an entity linker about people in Irish cultural heritage collections was developed from the Dictionary of Irish Biography, a compendium of biographies about notable Irish figures written by Irish historians. We present the design considerations which influenced the structure of the knowledge base given its intended application and the process by which the desired information was extracted from the collection of biographies. This includes the description of an automatic linking process for associating entities in the new knowledge base with their corresponding entries in DBpedia where such a correspondence was possible. We discuss other resources which possess similar properties to the Dictionary of Irish Biography that may be useful for expanding the current knowledge base. A preliminary test on a collection of Irish manuscripts is conducted using an existing Entity Linking tool to demonstrate the use of this knowledge base in a linking task.

Keywords: Digital Humanities; Knowledge Base Creation; Entity Linking; Cultural Heritage

1. Introduction

It is often said that one of the goals of research in Digital Humanities (DH) is to “open up” and “expose” cultural treasures so that they may be investigated and explored by anyone who wishes to engage with their contents [32]. Efforts to pursue this ambition often incorporate some level of entity based search, discovery and/or personalisation [1, 13, 26, 28]. While these tools are not always semantic in nature, there is certainly interest among the DH community

E-mail address: gary.munnelly@adaptcentre.ie

10.1016/j.procs.2018.09.019

with regards to organising and structuring digital collections using Linked Open Data principles. However, as greater quantities of Cultural Heritage (CH) material are digitised, it becomes increasingly infeasible to manually annotate these collections. One method of automating this process is Entity Linking.

Entity Linking (EL) is a well established problem in the field of Natural Language Processing [4, 19, 22, 27, 29]. Given a set of ambiguous surface forms which refer to entities, an automated process is tasked with unambiguously identifying a corresponding set of referents. This is done by mapping the input surface forms to URIs obtained from a knowledge base. Under ideal circumstances this results in an efficient, cost-effective, means of annotating flat, raw textual content with intelligent, semantic annotations.

Unfortunately, previous efforts to apply EL to CH collections have found the process to be rather challenging [2, 21, 30]. This is often due to poor representation of collection entities in commonly employed knowledge bases. The most obvious solution would be to find a better knowledge base, or else build one if no suitable source of information can be found. However, the structure and contents of a knowledge base can have a drastic impact on the effectiveness of the EL system. Hence due care must be taken when pursuing either of these solutions.

This paper presents the process by which a new knowledge base was developed for Irish CH collections. It is motivated by our own experience with performing EL on archive material and the observed poor coverage with respect to the entities that are relevant to our overarching research. We describe the process by which the knowledge base was developed and the factors which influenced our design decisions. We demonstrate the use of this knowledge base with REDEN [3], an EL tool designed to link a single collection against multiple knowledge bases.

2. Background

2.1. Previous Work

Our previous work has focused on automating the semantic annotation of a collection of 17^{th} century depositions obtained from individuals who were affected by the 1641 Irish rebellion. An evaluation to assess the performance of El systems in this task showed that the greatest limiting factor was the lack of entities in popular knowledge bases [21]. Finding a single knowledge base which adequately covers this domain is not possible due to the variety and nature of the entities in question. People range in importance from common servants up to nobles, military leaders and monarchs. While notable historical figures may be represented in popular knowledge bases, less important individuals are more challenging to identify. Locations are similarly problematic as several no longer exist due to shifting borders, or are known by a different name in modern times. Compounding this is the inconsistent spelling in the text of the documents, as the English language was not standardised until some time in the mid-18^{th} century. Sourcing a suitable knowledge base for this collection is thus, a frustrating problem.

In order to remedy this lack of information we have looked to sources commonly employed by historians who research the history of 17^{th} century Ireland. Unfortunately, these sources are disparate and often not semantically structured (assuming they are structured at all). Yet they are potentially helpful as we have found references to several entities mentioned in the depositions.

Specifically we have identified three primary sources which may help to inform an EL system:

- The Down Survey: A national survey of land ownership in Ireland after the 1641 rebellion. This provides the names of land owners and geographic locations documented in the depositions.
- The Statute Staple: Documents transactions between individuals, and provides information about debts owed between various parties before the rebellion
- The Books of Survey and Distribution: Again a report of land ownership used for taxation purposes.

While these sources offer the most complete listing of entities pertinent to the depositions that we are aware of, they are limited from the perspective of EL as they are simply lists of names of people. There is little evidence to help distinguish between two individuals with the same name other than place of residence. In some instance individuals are only referenced by title. It may be possible to extract relationship information from the Statute Staple, but these efforts quickly fall into the task of record resolution which is a challenging problem.

From the perspective of geography, however, these sources are extremely useful as historians have GIS tagged each county, barony and townland with the longitude and latitude of their modern equivalents.

We have also identified two secondary sources of information:

- the Dictionary of Irish Biography (DIB)
- the Oxford Dictionary of National Biography (ODNB)

Both are collections of biographies about notable Irish and British people which are written and maintained by professional historians. Although they are not as complete with respect to the depositions as the three primary sources mentioned above, they are a much more reliable, richer source of information. Individual biographies can be used to build a profile of historical figures and links between biographies can indicate relationships as well as yielding variant surface forms by which an entity may be referenced.

Current efforts have focused predominantly on organising ODNB and DIB into a knowledge base. In this paper we discuss our efforts to construct a knowledge base using DIB, but the methods described translate with little modification to ODNB as well.

2.2. A Brief Overview of Entity Linking and the Role of the Knowledge Base

As mentioned in Section 1, EL is essentially a mapping challenge between raw text surface forms and URIs in a knowledge base. The process begins by identifying a set of candidate referents to which each surface form may be referring. Often this is achieved by indexing all entities in the knowledge base using Lucene and then executing the surface form as a query against the index. The candidate set for each surface form is comprised of the set of results found in the index. The EL algorithm eliminates those candidates that do not make sense according to some set of metrics until it arrives at what it believes to be the best mapping from surface forms to referents. A vast array of tools and methods for performing EL now exist, but most approaches can essentially be reduced to an attempt at computing two types of similarities [22]:

1. **Local Similarity:** Does the mapping make sense given what is known about the surface form and the candidate referent in question? This can be as simple as the textual similarity between the surface form and known labels for the referent [24] or as sophisticated as a contextual comparison between the text from which the surface form was extracted and known contexts for the candidate referent [36].
2. **Global Similarity:** Does this mapping make sense given what is known about all other surface forms and their respective candidates? This is often done using graph based measures although some variant approaches do exist [29, 31, 35].

Precisely how these similarities are computed depends on the structure and content of the chosen knowledge base. The roots of EL lie in a problem known as Wikification, wherein Wikipedia formed the knowledge base and surface forms were mapped to the URL of corresponding Wikipedia articles [5, 10, 17]. While EL is now almost universally performed with respect to semantic knowledge bases, the target vocabulary is often derived from Wikipedia e.g. DBpedia [16] or YAGO2 [12]. These are attractive targets for a number of reasons, not least of which is the immense number of entities they document. These knowledge bases provide lists of the various surface forms by which an entity may be referenced, which is helpful for the candidate retrieval process. The Wikipedia articles from which DBpedia and YAGO2 entities are derived provide long-form descriptions of entities which may be used to determine key phrases and terms that indicate common contexts for entities [36]. Hyperlinks between pages (and subsequently between entities) suggest relationships between candidates [11]. Wikipedia links have also been used to compute simple semantic similarity measures between candidates for the purposes of EL [34]. It is this richness of information contained in Wikipedia derived knowledge bases which has led to the diverse range of EL methods that we see today.

In the case of CH, the collections encountered can be so diverse in nature that finding a single knowledge base with suitable coverage for all entities may be unlikely. Where appropriate knowledge bases can be identified, the information they contain may be too sparse for an EL system to work effectively. When dealing with this problem, an approach by Brando et al. [3] presents the interesting possibility of linking with respect to multiple knowledge bases. This approach, designated REDEN, uses a specialised knowledge base that is targeted at the collection being annotated. Where corresponding entities exist in other knowledge bases, the specialised knowledge base indicates this via `owl:sameAs` and `skos:exactMatch` properties which target the alternative knowledge base. The specialised knowledge base is used to generate a set of candidate referents while the alternative knowledge base helps to provide supporting information that can compensate for sparsity in the data.

This is an interesting approach for a number of reasons, but it is particularly appealing from our perspective due to the disparity of the sources from which we are attempting to construct our own knowledge bases. While it may not be appropriate to combine multiple historical resources into a single knowledge base, an EL system which can follow relationships between knowledge bases may avail of information that is distributed across multiple sources. Whether or not this is beneficial in our context is the subject of investigation.

2.3. Deliberately Limiting the Knowledge Base

A useful feature for CH that is often notably absent from EL systems is the ability to limit the scope of a system's search either by geography or by time. In addition to supplying content for annotation, a user would also supply a maximum century, or a bounding polygon outside of which an entity should not be considered as a candidate referent. This human-in-the-loop intervention would eliminate obviously incorrect candidates according to a hard threshold, preventing them from leading the linking algorithm astray. However, the general approach to working with EL systems would suggest that this filtering process is something that should be performed when the knowledge base is initially indexed, rather than during the linking process itself.

This leads to an interesting perspective on the knowledge base as both a means to inform and control the linking process. In the case of our own efforts to build a knowledge base that can inform linkers about Irish cultural heritage, the fact that we draw information from cultural resources that describe the British Isles implicitly limits the scope of the linker to the geography of this region. We also endeavour to capture information about dates of birth and death for the purposes of imposing these human-in-the-loop limits.

3. Constructing the Knowledge Base

3.1. Extracting Information from Biographies

The Dictionary of Irish Biography (DIB) is a collection of biographies about notable Irish figures "from the earliest times to the year 2002". Currently published in nine volumes, the biographies are also available online[1] and are continuously updated as historians contribute new biographies. The current collection stands at approximately 9700 biographies. These biographies were harvested from the DIB website and stored locally. Each biography has an associated unique identifier that is applied by DIB which was retained and used to identify each entity in the resulting knowledge base. The title and content for each biography was extracted along with all outbound hyperlinks.

While the biographies are not strictly structured, there are a number of consistencies in their formatting which make it possible to extract useful information using simple regular expressions. The name of the individual who is the focus of the biography is usually given in the title with the surname given first followed by the forename. A comma separates these two fields. Alternative forenames and surnames appear in parenthesis in their respective name-parts. Nicknames are quoted. The name is then repeated in the same format at the start of the first sentence of the biography. This is followed by dates of birth and death in parenthesis. See below for an example:

"Butler (le Botiller, Pincerna), Theobald (c. 1223-1248)"

The process of extracting information from the biography begins by attempting to generate lists of surface forms which might refer to the entity.

The name string is extracted from the title and split on the first comma that does not occur between parentheses. This divides the name into a surname and forename part. Parentheses are initially collapsed and a gazetteer of honorifics and titles is applied in order to remove titles such as "Earl" or honorifics such as "Sir". The remaining strings are tokenised

[1] http://dib.cambridge.org/

on whitespace. The first name in the remaining forename part and the last name in the surname part are respectively set aside as the individual's forename and surname. This would give "Theobald Butler" for the above string. The honorifics, titles and bracketed names are then reintroduced. The comma separated values between parentheses are split on commas yielding alternative surnames and/or forenames. These are combined in all possible permutations to generate a list of surface forms by which the entity may be referenced.

Nicknames are identified as alternative names that occur between quotes. These too are included in the permutations of the name, but are also individually added to the list of surface forms associated with the entity. For example, given the name "Daniel Joseph ('Dan') Bradley", both "Dan Bradley" and "Dan" would be considered surface forms for this entity.

Although this approach would seem bespoke to this collection, it may not be as inflexible as it first appears. The structure of surname, forename with alternative names in parentheses and nicknames in quotes is a common one which translates to ODNB. We have also found that historians typically transcribe individuals in Statute Staple as well the Books of Survey and Distribution according to this structure.

Similarly, the date of birth for each individual was extracted using a regular expression. While this field is generally consistent, there is some noise due to different writing styles. Some historians only give a date of birth or date of death. Some dates are uncertain, in which case alternative dates are listed after forward slashes e.g. (1710/11? - 1790). Some historians use an "x" rather than a slash. Sometimes only an approximate century is given e.g. "mid to late 7th". Once these variations were identified, accounting for them was simply a matter of creating an appropriate regular expression. Capturing the fuzzy timespans that they describe, however, was an important consideration when choosing a vocabulary for the knowledge base. An approximate distribution of biographies with respect to century is given in Figure 1 according to the extracted birthdays.

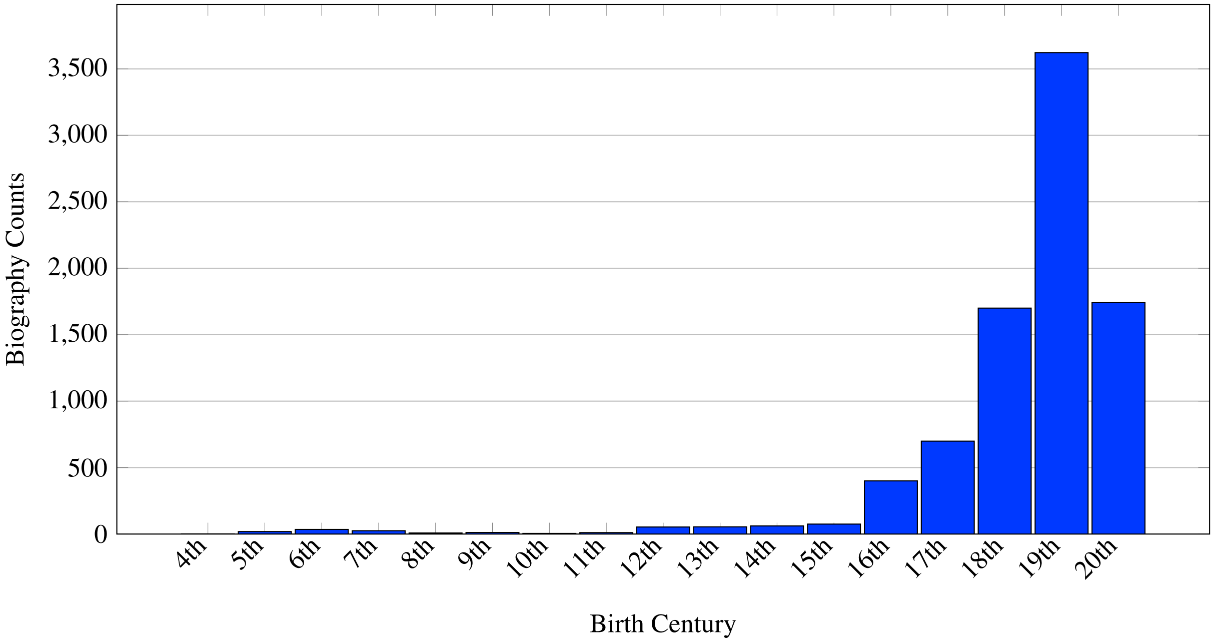

Fig. 1. Distribution of biographies by century in DIB

Relationships between entities are often an important feature employed by EL systems when discerning the referent for an entity mention. Hyperlinks between biographies in DIB were treated as directed relationships and added to the knowledge base as `dbo:related` properties. Initial tests also used the anchor text associated with a link as a means of extracting additional surface forms for the target entity. However this was found to introduce too much noise in the knowledge base and was ultimately removed.

The information extracted from the biographies was initially structured using the DBpedia ontology vocabulary for properties such as relationships between entities, dates of birth and death, and class of entity. FOAF was used to describe name parts extracted by the permutation process given above and to link back to the source biography for provenance via the `foaf:primaryTopic` property.

With respect to dates of birth, DBpedia's vocabulary is too limited to capture certain features of the biographies in DIB. For example, the concept of *florium* (Latin for "he/she flourished") is used to describe an approximate temporal span in which historians are aware that an individual was alive, but are unable to determine precise dates of birth and death. A more concrete demonstration may be found in the birth and death dates of the clergyman Charles Coote, which are are given as "(1712/13-1796)", meaning that his data of birth occurred somewhere between January 1st 1712 and December 31st 1713. To capture the fuzzy nature of these measurements, the CIDOC-CRM vocabulary was used [6] which permits modelling time spans as probability distributions. An example of the resource generated given the biography of Theobald le Botiller is given in figure 2.

```
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix : <http://example.com/dib/> .

:1248-01-01_1248-01-01 a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1248-01-01"^^xsd:date ;
    crm:P82b_end_of_the_end "1248-01-01"^^xsd:date .

:1223-01-01_1223-01-01 a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1223-01-01"^^xsd:date ;
    crm:P82b_end_of_the_end "1223-01-01"^^xsd:date .

:birth_a1292 a crm:E67_Birth ;
    crm:P4_has_time-span :1223-01-01_1223-01-01 ;
    crm:P98_brought_into_life dib:Theobald_Butler_a1292 .

:death_a1292 a crm:E69_Death ;
    crm:P100_was_death_of dib:Theobald_Butler_a1292 ;
    crm:P4_has_time-span :1248-01-01_1248-01-01 .

:Theobald_Butler_a1292 a dbo:Person,
            crm:E21_Person,
            foaf:Person ;
        rdfs:label "Theobald",
            "Theobald Butler",
            "Theobald Butler (le Botiller, Pincerna)",
            "Theobald Pincerna",
            "Theobald le Botiller" ;
        dbo:birthYear "1223-01-01"^^xsd:date ;
        dbo:deathYear "1248-01-01"^^xsd:date ;
        dbo:related dib:Richard_de_Burgh_a1131,
            :Theobald_Butler_a1291,
            :Theobald_Butler_a1293 ;
        dbo:sameAs <http://dbpedia.org/resource/Theobald_Butler,
            _3rd_Chief_Butler_of_Ireland> ;
        foaf:familyName "Butler" ;
        foaf:givenName "Theobald" ;
        foaf:name "Theobald Butler" ;
        foaf:primaryTopic "http://dib.cambridge.org/viewReadPage.do?articleId=a1292" .
```

Fig. 2. Sample RDF in Turtle format derived from "Butler (le Botiller, Pincerna), Theobald (c. 1223-1248)"

It is also worth mentioning that the first sentence of a biography usually states an individual's occupation, place of birth and/or residence and occasionally familial relations. It may be possible to extract this information using some simple pattern matching, entity recognition and/or part of speech tagging, however we found the resulting output of these approaches to be too unreliable and ultimately not useful for the linking process. Hence we opted not to use them.

3.2. Linking DIB to DBpedia

As highlighted in Section 2.2, methods of EL which can link with respect to multiple knowledge bases have been developed. However, these require explicit links between the respective knowledge bases on the basis of mutual entities. In order to create these links, a method of identifying common entities between DIB, ODNB and DBpedia was developed. This approach is effectively an EL method in itself, but has been used in this context to resolve knowledge bases. Below we document how this process was applied for mapping DIB entities to DBpedia entities.

During a pre-processing phase, the names, anchor text and content of the associated Wikipedia article for all DBpedia entities that belong to the class `dbo:Person` are indexed using an instance of the Solr[2] search engine. The title of each biography in DIB is then executed as a query against the search engine to retrieve a set of candidate DBpedia entities which may be a match. The top ten results from the search engine for each biography are chosen as candidates. A best matching DBpedia entity to which a DIB biography can be mapped is chosen from this pool of candidates according to the process described below.

Given a DIB biography $b \in \mathcal{B}$, and a set of up to ten candidates $\mathcal{P}_b$ the best matching DBpedia referent $p_b^* \in \mathcal{P}_b$ for a given biography is the one that maximises the expression:

$$p_b^* = \underset{p}{\operatorname{argmax}}\, \Psi(b, p), \forall\, p \in \mathcal{P}_b \tag{1}$$

Where $\Psi(b, p)$ is a similarity function. For our purposes this was computed based on the similarity between the title of the biography and the name of the entity in DBpedia, and the Word Mover Distance (WMD) [15] between the content of the biography and the content of the Wikipedia article that corresponds to the DBpedia entity.

The similarity between the biography title and the entity's name Φ is computed according to the Monge-Elkan Method [20]. The biography title b_{title} and name of a candidate p_{name} are lower-cased, tokenised and stopworded producing two sets of tokens $\mathcal{T}_b$ and $\mathcal{T}_p$. A bipartite graph is constructed from these tokens and edge weights are computed with Jaro-Winkler similarity [33]. Using Edmond's blossom algorithm [8] an optimal mapping $\mathcal{T}_b \mapsto \mathcal{T}_p$ is found giving $\mathcal{W}$, the set of weighted edges which comprise the mapping. Name similarity is the generalised mean of the edge weights in $\mathcal{W}$ as described by Jimenez et al. [14] with $m = 2$ for our purposes:

$$\Phi(b, p) = \left(\frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} w^m \right)^{\frac{1}{m}} \tag{2}$$

Content similarity Ω between the biography $b_{content}$ and the candidate's Wikipedia article $p_{article}$ is computed using gensim's [23] implementation of WMD Similarity (which is simply the negation of WMD). The word embeddings used by WMD are obtained from a Word2Vec model [18] trained on a Wikipedia dump excluding redirects, disambiguation pages etc.

The final value of Ψ was computed as a linear combination of these two functions with α and β being used to control how much influence either function could exert on the result:

$$\Psi(b, p) = \alpha\, \Phi(b_{title}, p_{name}) + \beta\, \Omega(b_{content}, p_{article}) \tag{3}$$

[2] http://lucene.apache.org/solr/

A hard threshold τ is applied to p_b^*, enforcing a minimum similarity between a biography and its final DBpedia equivalent $\overline{p}_b^*$ with NIL indicating that a biography does not have a DBpedia counterpart:

$$\overline{p}_b^* = \begin{cases} p_b^*, & \text{if } \Psi(b, p_b^*) > \tau \\ NIL, & \text{otherwise} \end{cases} \tag{4}$$

3.3. Evaluating Linking Quality

An evaluation to assess the accuracy of this method for linking biographies to DBpedia was performed using a sample of 200 biographies from DIB and 200 biographies from ODNB. For each biography in the sample dataset, a human annotator identified a corresponding referent in DBpedia where such a referent existed. In the event that no suitable DBpedia referent was found, the human annotator assigned the label "NIL" to the biography.

Because we can classify this method of linking biographies to DBpedia as an EL problem, we measured the accuracy of the linking process using the BAT framework [7] which computes the performance of an EL system according to Precision (P), Recall (R) and F-measure (F1). BAT defines two different methods of computing these three scores – micro and macro P, R and F1.

- P_{micro} and R_{micro} are obtained by treating the entire collection as a single large linking problem and computing the performance of the EL system with respect to all annotations in the evaluation corpus.
- P_{macro} and R_{macro} are found by computing P and R for each individual document in the evaluation corpus and averaging the results.
- Both $F1_{macro}$ and $F1_{micro}$ are the harmonic mean of the corresponding P and R values.

We note, however, that the process of evaluating an EL system for our specific problem can be drastically simplified. First, because the surface forms are known in advance and given to the EL system, P, R and F1 will always be exactly equal. Furthermore, because each "document" contains only a single entity (the subject of the biography), it can be seen that there is no difference between computing micro and macro P, R and F1. Hence, the computation of accuracy for this method when linking DIB and ODNB to DBpedia is simply:

$$\text{Accuracy} = \frac{\text{Correct Annotations}}{\text{Total Annotations}} \tag{5}$$

We found that a threshold $\tau = 0.55$ yielded the best performance with $\alpha = 0.1$ and $\beta = 0.9$. However, this is seen to be quite sensitive as illustrated in Figure 3. A slight variation in the value of τ has a large effect on the accuracy of the resulting annotations. It is reassuring that this threshold is consistent for both collections, but troubling that it is quite so exact in this context.

The decision to weight heavily in favour of content similarity rather than surface form similarity is due to the fact that the title often does little to discern one entity from another in the context of DIB. For example, in DIB there are six biographies about men named "Charles Coote". There is little evidence from the title to discern one Coote from another. However, an examination of the biography content shows that some were members of the clergy, others were soldiers and most were members of the nobility. Hence by comparing the content of the biography with the content of a Wikipedia article we can obtain a more reliable measure of similarity than if we were to weight in favour of surface form similarity. Yet we cannot completely disregard surface form similarity due to the possibility that the search engine may return a result for "Richard Coote" (or indeed some other Coote) when searching for "Charles Coote". A slight weight in favour of surface form similarity helps to reduce the effects of this noise.

In the initial evaluation we observed a large difference in performance between mapping DIB to DBpedia (81.5% accurate) versus linking ODNB to DBpedia (67.5% accurate) when $\tau = 0.55$. This was found to be due to the fact that several biographies obtained from ODNB contained no textual content. Instead, these biographies were comprised of pictures of the individuals in question. After filtering the contents of the ODNB dataset to remove any biographies which contained fewer than 100 words, the performance of our linking method for ODNB increased to 77.5%. Some of the remaining difference in performance can be ascribed to the fact that certain "biographies" in ODNB act as disambiguation pages, containing links to multiple different individuals who have the same name.

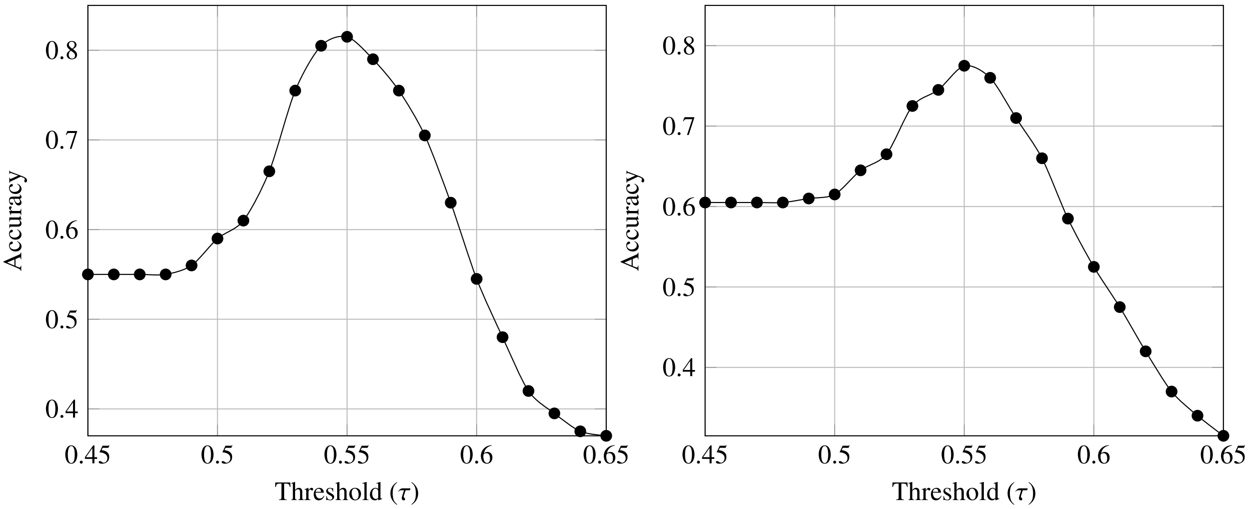

Fig. 3. Plots showing the performance of WMD Similarity on DIB (left) and ODNB (right) as the value of threshold τ varies from 0.45 to 0.65.

The relative performance of this linking method was examined by performing a similar evaluation with DBpedia Spotlight. Using the same gold standard corpus of 200 DIB biographies and 200 ODNB biographies, the biography content and the surface form of the entity in question were uploaded to the `disambiguate` API endpoint for DBpedia Spotlight[3]. As with our own linking method, the service was tasked only with identifying the corresponding DBpedia referent for an input biography. The surface form that identified the subject of the biography was provided to the El service as part of the request, reducing the margin for error that might be caused by spotting the wrong surface form. Under these conditions, DBpedia Spotlight was found to have a maximum accuracy of 55.5% with respect to DIB and 46.5% with respect to ODNB when the confidence threshold was set to 0.65 in the request. However, DBpedia Spotlight's accuracy was found to be considerably more stable than WMD Similarity with a minimum score of 52% for DIB and 45.5% for ODNB when the confidence threshold varied from 0 to 1.

Annotator	DIB	ODNB
WMD Similarity	**0.815**	**0.675**
WMD Similarity (Biography Length Filtered)	**0.815**	0.775
DBpedia Spotlight (Biography Length Filtered)	0.555	0.465

Table 1. Micro F1 Scores returned by BAT framework

4. Demonstration of Application

The resulting knowledge base derived from DIB and linked to DBpedia was used to create a knowledge base for REDEN. This is a reasonably simple task as REDEN uses a TSV file which maps surface forms to URIs in order to build a Lucene index for candidate retrieval. RDF files which describe the entities themselves and their respective relationships are harvested either from the web or from a local folder during the linking process. The necessary files

[3] http://model.dbpedia-spotlight.org/en/disambiguate

were generated and stored locally for DIB and REDEN was configured to run in online mode so that it could harvest additional information from DBpedia.

REDEN was applied to the same evaluation dataset which we used for our evaluation in [21], although it is not possible to assess the performance of REDEN based on this dataset as it is not yet annotated with URIs from the DIB knowledge base. Even so, by observing the output from REDEN we noticed that, in spite of the information acquired from DIB, REDEN struggled to identify the correct referent in the knowledge base. In some cases this would appear to be due to the challenging spelling of the depositions. This problem was expected and the simple solution is to introduce a fuzzy string matching process to the candidate retrieval phase. The most appropriate methods might be those used by researchers investigating record linkage in census data [9, 25].

However, REDEN also failed to annotate historical figures such as Sir Charles Coote [4] which we found surprising. The correct referent was included in the candidate referent pool during the linking process suggesting that the problem is the linking method itself. REDEN uses a simple graph measure based on the degree of nodes to assess the quality of a candidate. This is likely the source of the problem, but an investigation into the most appropriate solution is still underway.

5. Conclusion

The approach described in this paper has enabled us to produce an Irish CH knowledge base derived from DIB and linked to DBpedia. These same methods can be applied to ODNB, which is comprised of more than 75,000 biographies which are of similar structure to DIB in order to expand on the knowledge available to the EL system. Yet our most complete source of data is likely to be the three primary sources listed in Section 2.1. Overcoming the sparsity of information in these resources, however, will be challenging. It is possible that linking with respect to multiple knowledge bases will be beneficial, but there are clearly considerations beyond simply merging sources of information.

We have largely been able to automate the harvesting and structuring of data from DIB by exploiting patterns in how the documents are formatted. The process of establishing corresponding DBpedia articles is reasonably accurate, although it is clear that the data will require manual cleaning as errors come to light. Nevertheless, this has been a cost-effective, minimal effort means of acquiring data for linking in CH where traditional resources are found to be lacking.

Loosely underpinning this pursuit is how scholars will respond to annotations supplied by an EL service. Wikipedia is a somewhat controversial source of information for scholars, with experts often questioning the accuracy of its content. A debate about whether or not such suspicions are merited is beyond the scope of this paper. However, there is almost certainly something to be said for the development of knowledge bases which are built on sources used by expert scholars.

The most apt summary we have found to describe the solution we are pursuing is, “there is no silver bullet”. While this is true, it does not mean that EL is a hopeless endeavour wherever popular knowledge bases prove to be inadequate. Information about the entities in these collections exists in a variety of forms, although it may not be semantic. In extreme cases, the data may not even be structured. Yet an informed approach to organising this data may help to facilitate a more reliable EL process.

Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

[4] http://dib.cambridge.org/viewReadPage.do?articleId=a2018

References

[1] Agirre, E., Aletras, N., Clough, P., Fernando, S., Goodale, P., Hall, M., Soroa, A., Stevenson, M., 2013. Paths: A system for accessing cultural heritage collections, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 151–156.

[2] Agirre, E., Barrena, A., Lacalle, O.L.D., Soroa, A., Fern, S., Stevenson, M., 2012. Matching Cultural Heritage items to Wikipedia.

[3] Brando, C., Frontini, F., Ganascia, J.G., 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. Complex Systems Informatics and Modeling Quarterly , 60 – 80.

[4] Bunescu, R.C., Pasca, M., 2006. Using encyclopedic knowledge for named entity disambiguation., in: Eacl, pp. 9–16.

[5] Cheng, X., Roth, D., 2013. Relational inference for wikification. Urbana 51, 16–58.

[6] Cidoc, C., 2003. The cidoc conceptual reference model.

[7] Cornolti, M., Ferragina, P., Ciaramita, M., 2013. A framework for benchmarking entity-annotation systems, in: Proceedings of the 22nd international conference on World Wide Web, ACM. pp. 249–260.

[8] Edmonds, J., 1965. Paths, trees, and flowers. Canadian Journal of mathematics 17, 449–467.

[9] Efremova, J., 2016. Mining social structures from genealogical data. Ph.D. thesis. Technische Universiteit Eindhoven.

[10] Fernando, S., Stevenson, M., 2012. 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, Stroudsburg, PA, USA. pp. 101–106.

[11] Ganea, O.E., Ganea, M., Lucchi, A., Eickhoff, C., Hofmann, T., 2016. Probabilistic bag-of-hyperlinks model for entity linking, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee. pp. 927–938.

[12] Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G., 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence 194, 28–61.

[13] Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E., 2016. Warsampo data service and semantic portal for publishing linked open data about thesecond world war history, in: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (Eds.), The Semantic Web. Latest Advances and New Domains, Springer International Publishing, Cham. pp. 758–773.

[14] Jimenez, S., Becerra, C., Gelbukh, A., Gonzalez, F., 2009. Generalized mongue-elkan method for approximate text string comparison, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer. pp. 559–570.

[15] Kusner, M., Sun, Y., Kolkin, N., Weinberger, K., 2015. From word embeddings to document distances, in: International Conference on Machine Learning, pp. 957–966.

[16] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al., 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web 6, 167–195.

[17] Mihalcea, R., Csomai, A., 2007. Wikify!: linking documents to encyclopedic knowledge, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM. pp. 233–242.

[18] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

[19] Milne, D., Witten, I.H., 2008. Learning to Link with Wikipedia, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, ACM, New York, NY, USA. pp. 509–518.

[20] Monge, A., Elkan, C., 1996. The field matching problem: Algorithms and applications, in: In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 267–270.

[21] Munnelly, G., Lawless, S., In Press. Investigating entity linking in early english legal documents, in: Digital Libraries (JCDL), ACM/IEEE Joint Conference on.

[22] Ratinov, L., Roth, D., Downey, D., Anderson, M., 2011. Local and global algorithms for disambiguation to wikipedia, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics. pp. 1375–1384.

[23] Řehůřek, R., Sojka, P., 2010. Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta. pp. 45–50. http://is.muni.cz/publication/884893/en.

[24] Sasaki, F., Gornostay, T., Dojchinovski, M., Osella, M., Mannens, E., Stoitsis, G., Ritchie, P., Declerck, T., Koidl, K., . Introducing FREME: Deploying linguistic linked data., in: MSW@ ESWC, pp. 59–66.

[25] Schraagen, M.P., et al., 2014. Aspects of record linkage. Ph.D. thesis. Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University.

[26] Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B., 2008. Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. Web Semantics: Science, Services and Agents on the World Wide Web 6, 243 – 249. URL: http://www.sciencedirect.com/science/article/pii/S1570826808000620, doi:https://doi.org/10.1016/j.websem.2008.08.001. semantic Web Challenge 2006/2007.

[27] Shen, W., Wang, J., Han, J., 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Transactions on Knowledge and Data Engineering 27, 443–460.

[28] Steiner, C.M., Agosti, M., Sweetnam, M.S., Hillemann, E.C., Orio, N., Ponchia, C., Hampson, C., Munnelly, G., Nussbaumer, A., Albert, D., et al., 2014. Evaluating a digital humanities research environment: the cultura approach. International Journal on Digital Libraries 15, 53–70.

[29] Usbeck, R., Ngomo, A.C.N., Röder, M., Gerber, D., Coelho, S.A., Auer, S., Both, A., 2014. Agdistis-graph-based disambiguation of named entities using linked data, in: International Semantic Web Conference, Springer. pp. 457–471.

[30] Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R., 2015. Exploring entity recognition and disambiguation for cultural

heritage collections. Digital Scholarship in the Humanities 30, 262–279.

[31] Waitelonis, J., Sack, H., 2016. Named entity linking in# tweets with kea., in: # Microposts, pp. 61–63.

[32] Whitelaw, M., 2015. Generous Interfaces for Digital Cultural Collections. Digital Humanities Quarterly 9.

[33] Winkler, W., 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, in: Proceedings of the Section on Survey Research Methods, pp. 354–359.

[34] Witten, I., Milne, D., 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, in: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA, pp. 25–30.

[35] Yosef, M.A., Hoffart, J., Bordino, I., Spaniol, M., Weikum, G., 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. Proceedings of the VLDB Endowment 4, 1450–1453.

[36] Zwicklbauer, S., Seifert, C., Granitzer, M., 2016. Robust and Collective Entity Disambiguation through Semantic Embeddings, ACM Press. pp. 425–434.