Investigating Entity Linking in Early English Legal Documents

Gary Munnelly
Adapt Centre
Dublin, Ireland
gary.munnelly@adaptcentre.ie

Séamus Lawless
Adapt Centre
Dublin, Ireland
seamus.lawless@adaptcentre.ie

ABSTRACT

In this paper we investigate the accuracy and overall suitability of a variety of Entity Linking systems for the task of disambiguating entities in 17^{th} century depositions obtained during the 1641 Irish Rebellion. The depositions are extremely difficult for modern NLP tools to work with due to inconsistent spelling, use of language and archaic references. In order to assess the severity of difficulty faced by Entity Linking systems when working with these documents we use the depositions to create an evaluation corpus. This corpus is used as an input to the General Entity Annotator Benchmarking Framework, a standard benchmarking platform for entity annotation systems. Based on this corpus and the results obtained from General Entity Annotator Benchmarking Framework we observe that the accuracy of existing Entity Linking systems is lacking when applied to content like these depositions. This is due to a number of issues ranging from problems with existing state-of-the-art systems to poor representation of historic entities in modern knowledge bases. We discuss some interesting questions raised by this evaluation and put forward a plan for future work in order to learn more.

CCS CONCEPTS

• **Applied computing** → **Digital libraries and archives**; • **Information systems** → *Content analysis and feature selection*;

KEYWORDS

Named Entity Disambiguation, Digital Humanities, Cultural Heritage

ACM Reference Format:
Gary Munnelly and Séamus Lawless. 2018. Investigating Entity Linking in Early English Legal Documents. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3197026.3197055

1 INTRODUCTION

In this paper we present an evaluation of the performance of Entity Linking (EL) systems when applied to a collection of 17^{th} century depositions. The corpus is comprised of interviews with Irish citizens regarding alleged crimes committed against them during the 1641 Irish Rebellion. They are of interest to numerous parties for a variety of reasons, but the challenging nature of their content, discussed further in Section 3.1, makes them difficult for scholars to explore.

We wish to investigate how well state-of-the-art EL systems perform in the task of automatically linking spotted entities in the depositions with a suitable referent. This referent may be used to resolve multiple mentions of entities throughout the collection, or even to build links between disparate collections based on mutual entities. Ideally the application of EL would facilitate the imposition of a semantic structure on the depositions, allowing historians to execute more complex queries on the collection's content and enabling the provision of more sophisticated search, discovery and personalisation services.

This paper is concerned exclusively with the EL problem. While Named Entity Recognition (NER) is a related and similarly important task, given the challenging nature of the depositions' content we consider it to be a separate problem which requires a dedicated study if its own. The inconsistent nature of language in the depositions means that NER tools introduce too much noise for us to be able to investigate the positive and negative properties of the EL algorithms themselves.

We perform our investigation by first manually annotating a subset of the depositions with referent URIs taken from DBpedia. We observe that only a small percentage of mentions in the text can be linked with a suitable referent, demonstrating the severe penalty introduced to an EL system's performance if the referent knowledge base provides insufficient coverage for the chosen collection. This is a common problem observed when working with EL systems in Cultural Heritage (CH) [1, 23]. We evaluate the performance of various EL systems with respect to this ground truth using a standard benchmarking framework.

2 BACKGROUND

In this section we will provide a high level description of EL, what it is and how it works. This is not intended to be a thorough state-of-the art, nor is it a detailed tutorial on how to perform EL. For parties seeking more information we refer to the work of Shen et al. [17]. We will, however, discuss some previous efforts to employ EL in solving Digital Humanities (DH) problems.

2.1 A Brief Introduction to Entity Linking

Entity Linking (also called Named Entity Disambiguation) is a problem in computational linguistics whereby an automated process attempts to determine the specific subject of a reference to an entity found in free text. The input to an EL system is usually a series of entity mentions and a body of text from which the mentions were sourced. The program produces as output a corresponding

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06…$15.00
https://doi.org/10.1145/3197026.3197055

list of referents for each mention in the input set. For example, given the input text, "*The said deponent then fled to the County of Dublin*" and the entity mention "*County of Dublin*" extracted from the text, an EL system might return a DBpedia URI which identifies http://dbpedia.org/page/County_Dublin as the referent entity. The set of URIs which identify referents is obtained from a knowledge base that is part of the EL system. This knowledge base is usually derived from popular linked data repositories such as DBpedia or YAGO.

Given a suitable knowledge base, the EL system identifies a number of candidate entities to which the mentions in the source text might be referring. Taking again the example of Dublin, we could be referring to a city in Ireland, a community in America, a village in Belarus or something else[1]. It is the task of the EL system to look at the evidence available to it and establish which entities in its knowledge base are the most suitable referents for the mentions it received as input. If no suitable referent can be established then the system may label the mention as NIL, meaning it could not identify an appropriate referent.

There has been much research into the development of EL systems with methods ranging from simple string lookups [15] to more sophisticated methods which perform a lexical comparison between the context of a mention and a source text which describes an entity [27]. Many methods also make the assumption that entities which are mentioned in the same context are likely related by some common theme [21, 26]. Therefore the system can combine the evidence from multiple entities to establish sensible referents.

EL systems face a number of challenges when applied to CH collections. One of the most immediately identifiable problems is the quality of the knowledge base.

If an entity does not exist in the knowledge base then the EL system cannot know about it. Given the highly specialised nature of many CH collections, this is a serious problem as a large proportion of entities that are important to the collection itself are either poorly represented or even entirely omitted from knowledge bases that are based on DBpedia or similar Wikipedia derived resources. An ongoing challenge is to deal with the problem of Emerging Entities (EE), those entities which might appear multiple times in text but do not have a referent in the knowledge base. In the absence of a solution to the EE problem, methods of compensating for gaps in the knowledge base need to be established.

EL systems are also taxed by the prevalence of evolving entities in collections which span broad periods of time. For example, titles such as the "King of England" or the "Bishop of Meath" are passed from person to person as new people take on a particular role over time. It is extremely difficult to establish precisely which individual holds a given title based solely on contextual information derived from the content of a source text. Similarly it can be challenging to deal with entities whose names and titles change over time. A common example might be a soldier who receives a promotion. Such an individual may be referred to with the title "Lieutenant" in earlier texts, but "Captain" in later ones. Alternatively, a woman who marries will often change her family name to match her husband's. Capturing these evolving entities is extremely difficult as they are rarely well documented in the knowledge base.

[1] https://en.wikipedia.org/wiki/Dublin_(disambiguation)

With respect to the linking process itself, certain assumptions made by the linking algorithm are not upheld when applied to CH collections. For example, it is common to compare the context in which an entity mention is found with the context from which the candidate referent in the knowledge base was extracted. This assumes that both sources of information use language in the same manner. This is obviously a problem, as language is an evolving thing. When dealing with older collections, a contextual similarity measure based on co-occurring words, word embeddings or other similar measures is often an unreliable feature as the language of the knowledge base is usually obtained from more modern resources.

To provide an alternative example of problematic assumptions made by entity linkers, some systems make use of a candidate probability prior derived from the popularity of a candidate entity in the knowledge base (where "popularity" can be determined by a variety of different methods). The intuition is that the most popular referent for a given surface form will be the correct referent for the majority of instances of the surface form in text. For CH collections, popular entities are not necessarily good candidates and this prior can actually mislead the linking process by encouraging it to favour more popular modern entities over more sensible candidates that are relevant to the collection.

The range of challenges faced by EL systems when dealing with CH is broad. Careful consideration must be given to the nature of the collection, the methods employed by the EL system and the quality of information obtained from the knowledge base in order to ensure that the annotations provided by the entity linker are reliable.

2.2 Related Work

A number of interesting efforts have been made to investigate the applications of EL for DH problems.

Work by Van Hooland et al. [23] attempted to assess the suitability of NER and EL tools for use in DH. They experimented with three disambiguation services – Alchemy API, DBpedia Spotlight and Zemanta. They raised some interesting points regarding what exactly *is* the correct URI for an entity in any given context. This is an extremely important question, particularly when trying to disambiguate entities through the lens of history. To take an example from our own research, if we see a reference to "Ireland" in a 17^{th} century document, is the most appropriate disambiguation the landmass that forms the island of Ireland, the Kingdom of Ireland (which is probably most appropriate for the time) or the Republic of Ireland (which is a more modern reference)? Ultimately Van Hooland et al. suggested that perhaps the "best" referent is the one on which the majority of annotation sources can agree. Nevertheless, he calls for caution, awareness and education on the part of those who would employ such tools to ensure that we are not too trusting of what the machine tells us.

Work by De Wilde also sought to investigate the usefulness of EL for digital archives [25]. He investigated the applications of EL on German and Dutch documents ranging from the early 19^{th} century to the mid-20^{th} century. His texts had been digitised through a method involving Optical Character Recognition (OCR), meaning there is likely to have been some noise in the resulting data. De Wilde used a simple disambiguation method based on dictionary

lookups and SPARQL queries. Where there was more than one possible referent for an entity, De Wilde chose the longest match. This approach was extremely simple but achieved impressive results which matched the state of the art. De Wilde was very enthusiastic about his results and planned to integrate the output from his EL software into the search interface for "Historische Kranten" project. He also suggested that the noise introduced by OCR might not have too severe an effect on the quality of EL.

One of the more considered efforts to address the challenge of EL in CH is by Carmen Brando, Francesca Frontini and Jean-Gabriel Ganascia [2]. Their work focused on the problem of poor entity coverage in common knowledge bases. They developed a method which can disambiguate with respect to multiple knowledge bases simultaneously. Their method allows for specialised knowledge bases such as BnF[2] to be integrated with more general sources such as DBpedia. The general knowledge base can compliment the specialised one by providing additional information which can be used by the linking process.

Given a set of entity mentions as input, REDEN begins by retrieving a set of candidate referents from an index built on one the knowledge bases. The knowledge base used for candidate retrieval should be the one that is most representative of the collection being linked. References to the same candidates are then retrieved from the supporting knowledge bases using *owl:sameAs* and *skos:exactMatch* properties. Entities retrieved from all knowledge bases are then fused into a single unified graph representation of each candidate referent. Once the fusion process is complete REDEN applies a graph centrality measure to determine the correct referent for each mention.

REDEN is an extremely interesting example of an attempt to perform EL in CH. It does not rely on language similarity as one of its features, as this is known to be unreliable in CH. Instead it focuses purely on the graph structure. It also provides mechanics for limiting what parts of the knowledge base are indexed so that only entities from a particular geographic region or time period may be considered for linking. This is often noted as potentially useful behaviour by those who have attempted to perform EL on CH collections [9].

3 CORPUS

In this section we introduce the 1641 depositions which form the basis of our evaluation corpus. We present some of the history behind the documents and explain why they present a challenge for Computer Scientists. We will also explain how the depositions were prepared for use as part of this paper's experiment. Those interested in learning more about the depositions are referred to the 1641 website[3] or the Cultura project[4].

3.1 The 1641 Depositions

The 1641 Depositions are a collection 8,000 depositions or witness statements, examinations and associated materials, amounting to 19,010 pages and bound in 31 volumes. They document the various losses, military actions, attacks and transgressions inflicted on numerous individuals during the 1641 Irish Rebellion. In spite of some controversy surrounding the accuracy of certain witness statements, the depositions provide an fascinating window into the lives of people in 17^{th} century Ireland.

Through a painstaking process which spanned a number of years the depositions have been digitised and annotated in TEI format preserving all aspects of the source manuscripts including the original spelling, deletions, margin notes etc. A team of scholars manually examined the depositions to extract references to locations and people whilst simultaneously tagging the documents with the nature of their contents (murder, theft etc). The result is an extremely data rich historical digital corpus.

Linguistically the depositions are challenging to work with as English was still a developing language in 1641. The documents are rife with features which make them difficult to interpret for a modern English speaker. Among the most striking of these features are the vast array of spelling inconsistencies and a severe lack of punctuation. Often a deposition is comprised of a continuous run-on sentence with the phrase "and further saith that" seemingly being substituted for a full-stop. The extract below from the *Examination of Elizabeth Williams* provides an example of these qualities:

> The rest of this deponents husbands goods Garrett mc Eohee and Donell mc Cabe kept & detained from him they being in the possession of them at the begining of the insurreccion And this Examinate further saith that she her husband together with their whole family was remoued into the Towne, where they had of their owne goods onely two steares and one Barrell of oates dureing the whole tyme of 17 weekes And further saith that on the seccond of January 1641 the Rebells came abroad into the Towne and tooke her husband (Mr William Williams) Mr Gabriell Williams (her brother in law) Mr Ithell Jones her sisters husband together with a Scotchman one Thomas Tran & hanged them all in a Barne in the backsyd of their lodgings where they were in prison, That day suffered besides these fower about fowerteen or fifteene whoe were all hanged or stabbed or both in the Towne

These peculiarities mean that the depositions have the capacity to confound some of the most basic off-the-shelf NLP tools including part-of-speech taggers, sentence chunkers and NER tools. Previous work by Mitankin et al. [12] tackled the problem of normalising spelling in the depositions with great success, while the Cultura project [19] also ambitiously attempted to provide a personalised search experience over the depositions with entity-based approaches being core to a number of services. Yet a suitable, automatic method of resolving and disambiguating multiple mentions of entities has not yet been found.

Performing EL on the depositions is challenging for a number of reasons. Setting aside the problem of language structure, the very nature of the entities themselves present a problem. The vast majority of people mentioned in the depositions are common folk who have no representation in popular knowledge bases like DBpedia. Even seemingly significant figures (e.g. Florence Fitzpatrick, who is accused of committing a number of atrocities in County Offaly) are often not present.

[2] http://www.bnf.fr/en/tools/a.welcome_to_the_bnf.html
[3] http://1641.tcd.ie
[4] http://cultura-project.eu

In many cases people of great significance are referred to by title rather than by name, e.g. the "kinge of Spaine". This can be problematic as there is currently a king of Spain – Filipe VI – who, from the perspective of a naive disambiguation tool, is likely a much better referent than our intended target – Philip IV,

It is also worth noting (although we do not consider this problem in this paper) that some entities are referenced by lineage rather than by name, e.g. "The son of Lord Mountgarret".

Locations also present an issue. Land borders have changed over time, meaning that some locations no longer exist (e.g. the Barony of Upper Ossory) or have been divided into new sub-regions e.g. Talbotstown is now split into upper and lower Talbotstown. This makes it hard to establish a suitable referent in modern knowledge bases. In some instances the appropriate action is simply to not annotate those locations if the modern equivalent is too different from the historic one.

Sometimes resolving an entity is difficult simply because of how different the historical spelling is from the modern one e.g. "Barony of Fassadinin" has been transcribed as "Barrony of ffassa and Dyninge" in the depositions.

Hence performing any sort of automatic analysis on a collection like the depositions is extremely difficult for a variety of reasons. Considering EL in isolation is challenging enough in this context largely due to problems with popular knowledge bases and the under-representation of the entities in which we are interested.

3.2 Corpus Preparation

From the complete collection of depositions we sampled 16 documents to use for our evaluation. We chose documents that were approximately 800 words in length as we felt this would provide enough content per deposition that they would be interesting yet not be too onerous to annotate. Depositions were chosen randomly from geographically distributed counties across Ireland.

While the purpose of this corpus in the context of this experiment is to assess how well EL works, in future we would like to expand it for use in assessing a more complete pipeline including NER and some of the more fundamental NLP tools (tokenizers, chunkers etc).

To help with this, we performed some basic pre-processing that we would expect an appropriate library to perform in practice. We removed content from the files that was contained in the margins or that had been crossed out by the original scribe (these were marked by `<note>` and `<del>` tags in the original TEI files). We also broke the depositions into approximate sentences as, again, this is an operation that we would expect a suitably implemented tool to perform.

Using WebAnno [5], a human annotator read the selected depositions and attached a DBpedia URI to each identified entity. The focus was on locations and people. Where no suitable URI could be identified, the entities were given an appropriate NIL label:

`http://aksw.org/notInWiki/<entity_text>`

where `entity_text` was the surface form of the entity with spaces removed. This format conformed with the annotation suggested by the GERBIL wiki[5].

The annotated corpus contains 480 annotated instances of people and locations. These were found to refer to 283 unique entities of which only 64 were found to have a suitable referent in DBpedia. The remaining 219 were assigned a suitable NIL label.

[5]https://github.com/dice-group/gerbil/wiki/URI-matching#consequences

4 EXPERIMENT

Given the human-annotated depositions, the purpose of the experiment was to assess how well annotations provided by an EL system would match those of a human annotator. In order to perform this comparison we made use of the General Entity Annotator Benchmarking Framework (GERBIL) [22] as an experimentation platform.

GERBIL was developed to provide a simple, consistent, reproducible means of assessing the performance of EL systems on different datasets. Users of the platform can configure an experiment by selecting a set of EL systems, an evaluation dataset and an evaluation method. Gerbil executes the experiment under the given conditions and returns the results in tabulated format. We will discuss the metrics by which these results are compared in Section 5.

As new EL systems are developed, their creators can register their API with GERBIL so that their technology may be used in future experiments. At the time of writing the platform has 17 registered annotation systems and 32 evaluation datasets.

The experiment configuration interface also allows users to upload custom datasets in NLP Interchange Format (NIF) as well as providing a hook for custom implementations of EL systems.

In our setup we chose to use *Disambiguate to Knowledge Base* (D2KB) as our experiment type. Under this configuration the EL systems are provided with the source text of each deposition and the already extracted entities. The only task which the EL systems need to perform is the assignment of URIs to each mention. This simplifies the experiment as the EL systems do not need to perform NER on the source text. We chose to run the experiment in this manner because our interest is in the ability of the system to accurately identify entities, rather than its ability to process the challenging language of the depositions. Resolving unconventional or archaic entity references to a modern referent is challenging enough.

It is also worth noting that some of the EL systems provided by GERBIL cannot perform NER and would need be omitted from experiment if NER was a requirement.

The depositions described in Section 3 were uploaded as a custom dataset to GERBIL and the experiment was configured to evaluate all available annotation systems against the collection. Under these conditions GERBIL ran the experiment.

5 RESULTS

Under our experiment configuration, GERBIL returns a vast array of statistics which must be interpreted. We have organised and presented these results across Tables 1, 2, 3, 4 and 5. Of the 17 annotation systems tested, 9 failed to finish due to internal errors. Problems like this usually occur because the service is offline and cannot respond to the experiment requests. Hence we have only reported the statistics from the 8 annotators which successfully completed annotating the depositions [3, 7, 10, 13, 16, 18, 21, 24]. A brief summary of the methods employed by the EL systems which successfully completed the task is given below, followed by an

Table 1: Results of D2KB evaluation obtained from GERBIL

Annotator	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
AGDISTIS	**0.5979**	0.5979	**0.5979**	**0.6052**	0.6052	**0.6052**
Babelfy	0.1299	0.2941	0.0833	0.1130	0.3348	0.0743
DBpedia Spotlight	0.1449	0.4767	0.0854	0.1281	0.4970	0.0774
Dexter	0.1082	0.3333	0.0646	0.0933	0.3536	0.0580
FOX	0.4051	**0.6327**	0.2979	0.4054	**0.6791**	0.2999
FREME NER	0.1012	0.3118	0.0604	0.1045	0.3076	0.0694
Kea	0.1466	0.3358	0.0938	0.1363	0.3384	0.0923
PBOH	0.4250	0.4250	0.4250	0.4266	0.4266	0.4266

Table 2: Results of D2KB evaluation obtained from GERBIL considering InKB

Annotator	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
AGDISTIS	**0.3395**	0.4589	0.3063	**0.3557**	0.4040	0.3177
Babelfy	0.2229	0.3348	0.1858	0.2439	0.2941	0.2083
DBpedia Spotlight	0.2667	**0.4970**	0.1959	0.2950	**0.4767**	0.2135
Dexter	0.1865	0.3536	0.1444	0.2175	0.3333	0.1615
FOX	0.3189	0.5176	0.2604	0.3077	0.4000	0.2500
FREME NER	0.2025	0.3076	0.1837	0.2035	0.3118	0.1510
Kea	0.2518	0.3384	0.2373	0.2761	0.3358	0.2344
PBOH	0.2696	0.2203	**0.3834**	0.2799	0.2292	**0.3594**

Table 3: Results of D2KB evaluation obtained from GERBIL considering EE

Annotator	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
AGDISTIS	**0.7189**	0.6858	**0.7840**	**0.7326**	0.6869	**0.7847**
Babelfy	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DBpedia Spotlight	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dexter	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FOX	0.4557	**0.9050**	0.3261	0.4822	**0.8962**	0.3299
FREME NER	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Kea	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PBOH	0.5565	0.7561	0.4612	0.5782	0.7542	0.4688

Table 4: Results of D2KB evaluation obtained from GERBIL considering GSInKB

Annotator	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
AGDISTIS	0.3063	0.3063	0.3063	0.3177	0.3177	0.3177
Babelfy	0.2571	0.5373	0.1779	0.2879	0.5278	0.1979
DBpedia Spotlight	0.3026	**0.7382**	0.1959	0.3361	**0.7885**	0.2135
Dexter	0.2096	0.4900	0.1444	0.2480	0.5345	0.1615
FOX	0.3743	0.7215	0.2604	0.3569	0.6234	0.2500
FREME NER	0.2469	0.4306	0.1837	0.2397	0.5800	0.1510
Kea	0.3042	0.6085	0.2164	0.3281	0.6562	0.2188
PBOH	**0.3834**	0.3834	**0.3834**	**0.3594**	0.3594	**0.3594**

Table 5: Performance statistics and configurations

Annotator	Errors	avg millis/doc	Threshold
AGDISTIS	0	6,049.6875	0.0000
Babelfy	0	4,288.4375	0.0607
DBpedia Spotlight	0	155.3125	0.0000
Dexter	0	2,734.1250	0.0000
FOX	0	23,157.5625	0.0000
FREME NER	0	347.5625	0.0000
Kea	1	9,240.3333	0.4833
PBOH	0	1,930.6875	0.0000

explanation of how GERBIL computes the values in each of the tables.

5.1 Entity Linking Systems

AGDISTIS is a graph based EL system which uses the well known HITS algorithm to select referent entities [21]. A set of candidates are retrieved from the knowledge base and a breadth first search is executed on candidate outbound links in order to construct a graph. HITS is executed on the graph and the candidate with the highest combined authority/hub score for each mention is selected as the referent.

Babelfy combines the tasks of Word Sense Disambiguation (WSD) and EL in order to present a unified method of semantically annotating text. During a pre-processing stage, a set of semantic signatures are generated for all concepts present in the knowledge base. After these signatures have been generated, an arbitrary input text may be processed for linking. Candidates for all mentions in the text (both entities and words) are retrieved and a graph is constructed with edges being added between candidates which have similar semantic signatures. A dense subgraph is then computed to determine the appropriate referents for all input mentions.

DBpedia Spotlight uses a Vector Space Model (VSM) in order to choose an appropriate referent for each surface form [10]. Every entity in the knowledge base is assigned a contextual description comprised of the concatenation of all paragraphs that reference the entity. DBpedia Spotlight also weights terms in this contextual aggregate according to how many entity contexts they are associated with. The EL process itself is essentially executed as an Information Retrieval problem. The similarity between an input mention (and the contextual text surrounding the mention) is compared with all of its candidate referents using cosine similarity. The candidate with the highest similarity score is chosen as the referent.

Dexter is an NER and linking framework which implements three different EL methods from the literature – TagME [6]: a vote based method which uses Wikpedia-Link Base Measure between candidates combined with the probability that a given anchor text points to a candidate entity; Collective Linking [8]: a graph based method which assigns weights to candidate entities based on a combination of importance of the mention to the surrounding context, compatibility of the candidate with the mention and coherence of the candidate with respect to other candidates; and WikiMiner [11]: a machine learning method which uses the probability that a surface form refers to a given candidate combined with the relatedness of the candidate to the surrounding context using Wikipedia-Link Based Measure.

FOX is actually a NER tool which incorporates EL as one of its outputs [18]. In order to perform EL, it uses its own deployment of AGDISTIS (described above).

FREME treats the problem of mapping surface forms to URIs as an Information Retrieval problem. The surface forms are executed as queries against a search index of entities which acts as the knowledge base. The top ranked entity for each surface form is chosen as the referent. The service also provides the option to re-rank the results from the search engine based on surface form similarity between the mention and the candidate referent's surface form, but this is not the default behaviour and is not part of this evaluation.

Kea implements a four stage EL process. At each stage in this four step process, if KEA believes it has found the correct referent for any given entity then it will commit to that referent and will not proceed to the next step.
From the set of candidates a graph is generated based on links between entities. Links are only created between candidates which are not competing directly with each other. First the algorithm considers connected components in the graph. The assumption is that the correct referents will form a long chain of connections. Next the algorithm checks to see how many of the candidates co-occur on each others' Wikipedia pages. After this a ranking algorithm such as PageRank or HITS is applied to find authoritative candidates. Lastly, if all else fails, a "negative context" step is applied which discards any candidates that do not fit with any referents that were chosen earlier in the disambiguation process [20].

Probabilistic Bag Of Hyperlinks (PBOH) learns a probability distribution based on the likelihood of a candidate being the correct candidate given the surface form by which it is referenced, the context obtained from the surrounding text and the joint probabilities of all candidates appearing together [7]. This problem is NP-hard, hence the resulting probabilities are approximated in practice using loopy belief propagation [14].

5.2 Evaluation Metrics

GERBIL is based on the BAT framework [4], which was designed to provide a consistent and fair means of assessing the relative performance of different EL systems. The BAT framework defines a number of different entity annotation problems on which an evaluation might be based, a vocabulary for describing EL systems and a set of metrics for assessing the output of said systems.

GERBIL adopts the evaluation metrics suggested by BAT, namely micro and macro precision (P), recall (R) and $F1$. Definitions for these are provided in the original paper [4], but we provide them below for clarity. In the context of EL P, R and $F1$ are computed using the number of true/false positives and true/false negatives. These may be defined as:

- True Positives (tp): The number of correctly annotated entities.
- True Negatives (tn): The number of correctly ignored entities (these values are not actually used, but we have included it for completeness).
- False Positives (fp): The number of entities which were annotated when they should have been ignored.
- False Negatives (fn): The number of entities which were ignored when they should have been annotated.

Given these definitions, the standard formulae for computing the values of P, R and $F1$ may be expressed as:

$$P = \frac{|tp|}{|tp| + |fp|}$$
$$R = \frac{|tp|}{|tp| + |fn|} \quad (1)$$
$$F1 = \frac{2 \times P \times R}{P + R}$$

Micro and macro P, R and $F1$ evaluate the annotators by taking two different perspectives on the collection.

Micro considers the entire collection as a single disambiguation problem. The total scores for tp, fp and fn are calculated across the entire collection and used to compute P_{micro}, R_{micro}. This, of course, lends greater weight to longer documents which are comprised of more entities. $F1_{micro}$ is then computed as the harmonic mean of P_{micro} and R_{micro}. The formulae for these values are given below:

$$P_{micro} = \frac{\sum_{d \in D} |tp_d|}{\sum_{d \in D} |tp_d| + |fp_d|}$$
$$R_{micro} = \frac{\sum_{d \in D} |tp_d|}{\sum_{d \in D} |tp_d| + |fn_d|} \quad (2)$$
$$F1_{micro} = \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}}$$

Macro treats each document as an individual disambiguation problem and then produces final evaluation scores by averaging the performance of the system on each document. In other words, P and R are calculated for each document using the formulae given in Equation 1. The values of P_{macro} and R_{macro} are the average of P and R scores obtained for each document. $F1_{macro}$ is then computed as the harmonic mean of P_{macro} and R_{macro}. The formulae for these values are given below:

$$P_{macro} = \frac{\sum_{d \in D} P_d}{|D|}$$
$$R_{macro} = \frac{\sum_{d \in D} R_d}{|D|} \quad (3)$$
$$F1_{macro} = \frac{2 \times P_{macro} \times R_{macro}}{P_{macro} + R_{macro}}$$

In the event of a division by zero in any of the Equations 1, 2, 3, GERBIL responds in one of two ways. If all tp, fp and fn values are zero, then P, R and $F1$ are assigned the value 1. Alternatively, if tp is zero but fp or fn are non-zero then P, R, and $F1$ are zero. This behaviour is documented on the GERBIL wiki[6].

Values for micro and macro P, R and $F1$ are computed for each annotator under four different experiment conditions, the results of which are displayed in Tables 1, 2, 3, 4.

Table 1 presents the results for a standard evaluation. All tp, fp and fn results are considered for all entities in the collection. This is an overall summary of how well each annotator performed.

Table 2 presents the results when we only consider responses from the annotators which are contained in the knowledge base (in this case, DBpedia). This essentially considers how well the annotator performed if we ignore Emerging Entities.

Table 3 presents the results of the experiment when only Emerging Entities (EE) are considered. Emerging Entities are entities that are not in the knowledge base. In the case of this experiment, anything which could be termed an Emerging Entities was given a NIL label in the gold standard. Hence this score can be considered a measure of how often an annotator correctly abstained from applying a label to a document entity.

Table 4 presents the results if we only consider entities in the gold standard that are present in the knowledge base. Again, this eliminates emerging entities, but we also only consider URIs returned by the annotators if the URI is applied to an entity in the gold standard whose correct annotation is contained in the knowledge base. Essentially if we only consider the entities that the annotator *should* have annotated and *did* annotate, then how many of those annotations were correct.

Table 5 is included to report some performance and configuration information for the annotators while the experiment was being run. Errors is a total count of errors reported by the annotation system. Avg millis/doc is the average number of milliseconds taken to annotate each document in the gold standard. Finally, some annotators have a confidence threshold. If the confidence of an annotator in its selected URI is below this threshold then the annotator will not apply the given URI to the corresponding mention. We had no control over the threshold value, but we wish to report it for the sake of completeness.

6 DISCUSSION

Looking at the results in Table 1, the annotator which seems to clearly stand out above the rest is AGDISTIS. However, a closer

[6]https://github.com/dice-group/gerbil/wiki/Precision,-Recall-and-F1-measure

examination of the results shows that it achieved the best performance in the EE task and actually performed quite poorly in the InKB task (albeit still better than the other annotators). This suggests that AGDISTIS' stellar performance is largely because it abstained from annotating most of the entities in the depositions. Because NILs comprise about 77% of the unique entity mentions, this was sufficient to increase its score immensely as demonstrated by the reported scores in Table 3. FOX (which is built on AGDISTIS) and PBOH also exhibit relatively appreciable overall performance for the same reason.

It is interesting that other annotators do not seem to succeed at abstaining and always apply a URI to a surface form. This, overzealous, approach to annotating entities may be extremely problematic if these systems were deployed in practice. Scholars who would use tools such as EL systems (or services built on top of the outputs they produce) need to know that they can trust the data with which they are being presented. If an annotation system is prone to annotating a collection with inaccurate links then it is of little use to history scholars.

The results of the InKB and GSInKB evaluation show that no single annotator is particularly suited to identifying the correct referent where one exists, although AGDISTIS does still perform better than the other EL systems considered. DBpedia Spotlight's performance warrants note as its EL method is based on comparing the context of the mention with known entity contexts in DBpedia using a VSM. This approach is obviously hobbled by the inconsistent language of the depositions, yet it still performs better than Babelfy, FREME, Kea and Dexter.

While none of the results for the annotation tasks on the InKB and GSInKB problems are especially good, certain graph based approaches appear to be more robust against the anomalies that we see in the depositions. AGDISTIS which uses HITS, and PBOH which builds probability distribution based on links between pages seem to perform relatively well when compared with other systems. This suggests that being able to exploit the relationships between candidates when selecting a referent may be an important consideration when choosing an EL system for collections such as the depositions. Yet Babelfy uses a dense subgraph approach while Kea uses connected components, neither of which appear to have resulted in particularly accurate annotations.

A statistic which is not reported by GERBIL is how often the correct referent was considered as a candidate and then ultimately rejected by the EL method, as opposed to the situation where the correct candidate was not identified as a potential candidate in the first place. It would be interesting to see how much of the inaccuracy of the surveyed systems can be ascribed to the candidate retrieval process as opposed to the EL method.

Clearly a large problem that we face is the lack of representation for the deposition entities in popular knowledge bases. Of the 283 unique entities which were manually annotated in the gold standard, only 64 (23%) were found to have a referent in DBpedia. One possible solution is to identify alternative, specialised sources of knowledge which can work in tandem with more common knowledge bases much like Brando's approach. An alternative (and likely more sustainable) approach would be to focus on the Emerging Entities problem. Given the ad hoc nature of the entities encountered in the depositions (often servants or soldiers), this probably makes more sense as identifying an all-encompassing knowledge base will be difficult.

We acknowledge that a fundamental weakness in our method is the lack of annotators involved in creating the evaluation corpus. Unfortunately, due to the specialised nature of the depositions, finding annotators with the prerequisite knowledge to annotate the entities is challenging. We are presently working with historians to improve the quality of the evaluation corpus.

Ultimately we would like to produce a corpus that is comprised of a more representative number of depositions, a minimum of 64: two documents for each county in Ireland. Alternatively it has been suggested that greater benefit may be derived from focusing on a specific county as this will help to constrain the problem that the corpus represents. However this seems like an unrealistic constraint as choice of county could have a drastic effect on the difficulty of the resulting corpus. This is because some counties are likely to have better representation in knowledge bases than others. For example, there are DBpedia entries for specific streets in Dublin City whilst some towns in the neighbouring County Meath have little-to-no representation.

7 CONCLUSIONS

Overall we believe that this research has raised some interesting questions about the properties of a "good" EL system for CH. There is still much work to be done. We will proceed by expanding and enhancing the gold standard 1641 depositions so that we can perform more rigorous evaluations. We will continue to work with historians to ensure its integrity.

We will also continue to seek answers to the problem of dealing with poor representations of entities in knowledge bases. This is undeniably our greatest challenge. The depositions provide many excellent examples of the obstacles faced when dealing with niche collections.

Finally we will work to discern the traits and qualities possessed by the tested annotation systems which resulted in their success (or otherwise) during this evaluation.

We believe that finding concrete answers to the questions raised by this paper will allow us to create powerful EL tools which will help to build meaningful links within and across archives.

ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 13/RC/2106) and the ADAPT Centre (www.adaptcentre.ie) at Trinity College, Dublin.

REFERENCES

[1] Eneko Agirre, Ander Barrena, Oier Lopez De Lacalle, Aitor Soroa, Samuel Fern, and Mark Stevenson. 2012. Matching Cultural Heritage items to Wikipedia. In *LREC*. 1729–1735.

[2] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly* 7 (July 2016), 60 – 80.

[3] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*. ACM, 17–20.

[4] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 249–260.

[5] Richard Eckart de Castilho, Eva Mujdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 76–84.

[6] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1625–1628.

[7] Octavian-Eugen Ganea, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann. 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, 927–938.

[8] Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 765–774.

[9] Hugo Manguinhas, Nuno Freire, Antoine Isaac, Juliane Stiller, Valentine Charles, Aitor Soroa, Rainer Simon, and Vladimir Alexiev. 2016. Exploring comparative evaluation of semantic enrichment tools for cultural heritage metadata. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 266–278.

[10] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*. ACM, 1–8.

[11] David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 509–518.

[12] Petar Mitankin, Stefan Gerdjikov, and Stoyan Mihov. 2014. An Approach to Unsupervised Historical Text Normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*. ACM, New York, NY, USA, 29–34. https://doi.org/10.1145/2595188.2595191

[13] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics* 2 (2014), 231–244.

[14] Kevin P Murphy, Yair Weiss, and Michael I Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 467–475.

[15] Felix Sasaki, Tatiana Gornostay, Milan Dojchinovski, Michele Osella, Erik Mannens, Giannis Stoitsis, and Phil Ritchie. 2015. Introducing FREME: Deploying Linguistic Linked Data.. In *MSW@ESWC*. 59–66.

[16] Felix Sasaki, Tatiana Gornostay, Milan Dojchinovski, Michele Osella, Erik Mannens, Giannis Stoitsis, Phil Ritchie, Thierry Declerck, and Kevin Koidl. [n. d.]. Introducing FREME: Deploying Linguistic Linked Data.. In *MSW@ ESWC* (2015). 59–66.

[17] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2015), 443–460.

[18] René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Named entity recognition using FOX. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*. CEUR-WS. org, 85–88.

[19] Christina M Steiner, Maristella Agosti, Mark S Sweetnam, Eva-C Hillemann, Nicola Orio, Chiara Ponchia, Cormac Hampson, Gary Munnelly, Alexander Nussbaumer, Dietrich Albert, et al. 2014. Evaluating a digital humanities research environment: the CULTURA approach. *International Journal on Digital Libraries* 15, 1 (2014), 53–70.

[20] Nadine Steinmetz and Harald Sack. 2013. Semantic multimedia information retrieval based on contextual descriptions. In *Extended Semantic Web Conference*. Springer, 382–396.

[21] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. AGDISTIS-graph-based disambiguation of named entities using linked data. In *International Semantic Web Conference*. Springer, 457–471.

[22] Ricardo Usbeck, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. 2015. GERBIL: general entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 1133–1143.

[23] Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities* 30, 2 (2015), 262–279.

[24] Jörg Waitelonis and Harald Sack. 2016. Named Entity Linking in# Tweets with KEA.. In *# Microposts*. 61–63.

[25] Max De Wilde. 2015. Improving Retrieval of Historical Content with Entity Linking. In *New Trends in Databases and Information Systems (Communications in Computer and Information Science)*, Tadeusz Morzy, Patrick Valduriez, and Ladjel Bellatreche (Eds.). Springer International Publishing, 498–504.

[26] Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment* 4, 12 (2011), 1450–1453.

[27] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016. Robust and Collective Entity Disambiguation through Semantic Embeddings. ACM Press, 425–434.