TRINITY COLLEGE DUBLIN

Entity Linking for Text Based Digital Cultural Heritage Collections

Author:
Gary MUNNELLY

Supervisor:
Prof. Séamus LAWLESS

A thesis submitted in fulfilment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

TRINITY COLLEGE DUBLIN

APRIL, 2020

Declaration

I, the undersigned, declare that this work has not been previously submitted as an exercise for a degree at this or any other University, and that, unless otherwise stated, it is entirely my own work.

Signature _____ Date _____

Gary Munnelly

Permission to lend or copy

I, the undersigned, agree that the Trinity College Library may lend or copy this thesis upon request.

Signature _____ Date _____

Gary Munnelly

Anything you absorb you will ultimately secrete. It's inevitable. Most of us are an original painting, and it's a mystery to us what is learned and what is borrowed, what is stolen and what is born, what you came with and what you found while you were here.

— Tom Waits

Acknowledgements

First and foremost I would like to extend my most profound and sincere gratitude towards my supervisor, Professor Séamus Lawless. I could not have finished this thesis without his guidance, support and (dare I say) tolerance for my interminably distractible nature. Professor Lawless gave me an opportunity to pursue and explore an ambition I have held since my childhood to the fullest extent of my own personal interests. For that I am eternally grateful.

I also wish to thank Professor Owen Conlan, who stepped in to help with organising and preparing for my viva voce after Shay unexpectedly passed away not long after this thesis was submitted.

My PhD has been guided by two postdoctoral researchers over the span of its life – Dr. Rami Ghorab during the early years and Dr. Annalina Caputo in the later. Thank you both for believing in my work and offering me support when I needed it.

On the spectrum of Digital Humanities, it is fair to say that my expertise is firmly rooted in the digital end of the scale. I would like to thank the multitude of humanities scholars in Trinity College who helped me to understand the other half of this whole. Dr. David Brown, Prof. Jennifer Edmond, Prof. Jane Ohlmyer, Dr. Micheál Ó Siochrú, Dr. Mark Sweetnam and Dr. Ciaran Wallace, your patience with this unruly engineer is greatly appreciated.

There are several people who have made this process infinitely more tolerable simply by virtue of their continued friendship in spite of my increasingly hostile sleep patterns, lack of sociability and rapidly dwindling grab-bag of conversational topics. In descending alphabetical order by surname (for the sake of speedier search and discovery) they are Nicole Basaraba, Caoilte Bashford, Emer Browne, Robert Browne, Natalie Duda, Michelle Feighney, Maurice Gavin, Cliodhna Gillen, Ciara Hackett, Emma Mahon-Smyth, Angie McLaughlin, Aaron Meehan, David Murphy, Sorcha Ní Cheallaigh, Harshvardhan Pandit, Madison Porter, Kate Ryan, Brendan Spillane, and Wes Shaw. These people kept me human. Sort of.

I promised the 2016 Alpha 2 group from the Trinity Walton Club that I would give them a mention. It was truly a joy to work with you all. I'm not sure what the future holds for you, but I am certain that it is bright.

Finally, I wish to thank my family: my parents, Mary and Nigel, and my siblings, David and Megan. Most people who have followed me on this journey have had the option to opt out whenever I became unbearable. More unfortunate individuals have been eternally tied to me by a cruel twist of fate and kinship. In many ways, the stresses and strains of my PhD were also theirs. Yes, I have eaten today. No, I'm not strapped for cash. I'll sleep when I'm tired. It *is* a real job.

Thank you all.

Dedication

This thesis is dedicated to the memory of Professor Séamus Lawless, who went missing hours after summiting Mount Everest in May 2019.

I have left my original acknowledgement to Shay intact, and unedited as it was on the day we first submitted this thesis in April 2019. It was intended to be a cheeky wink to Shay, acknowledging how supportive he was as a supervisor without stating outright how challenged I was by the whole process, and how much his support meant to me. That was a conversation I was fortunate to have with him before he left for Nepal.

The truth of the matter is, this PhD taxed and vexed me in ways I could never have anticipated, and it was not a challenge that I met well. At times Shay was as much a counsellor as he was a supervisor. His patience was truly infinite.

Shay was a wonderful educator, a good friend, and an exemplary human being. His loss is unquantifiable. In working with him I, of course, expanded my awareness of a specific field in computer science. But I also learned a great deal about compassion and consideration for people. It is these lessons in particular that I hope to carry forward.

Rest in peace Shay.

Abstract

The ongoing digitisation of cultural heritage data and subsequent publication of that data in digital format has completely changed the manner in which people investigate and engage with cultural treasures. This change has propagated from the interested user browsing the web to the expert scholar in a research setting, applying digital tools in their pursuit of answers to questions. As the rate at which content can be digitised increases and the scale of collections grows, there is an implicit need to provide accurate, automated methods of organising and structuring this content in meaningful ways.

The research presented in this thesis represents an investigation into the applications of Entity Linking techniques to the content of cultural heritage collections. A specific challenge faced in the context of this research is the specialised domain knowledge required on the part of the reader who must interpret these collections. It is here that contemporary sources of knowledge for the Entity Linking process are found to be lacking. Indeed, finding any individual source of information that can be used to adequately annotate this type of content is difficult. The challenge of identifying references to obscure entities is compounded by the extremely noisy nature of the content of these texts.

An investigation is performed into the state of existing Entity Linking solutions in order to identify approaches which may be robust to the challenges presented by this content type. Evaluations are run to test the efficacy of off-the-shelf Entity Linking solutions. This investigation demonstrates the severe difficulty faced by typical Entity Linking tools when dealing with this content type. An interesting approach is identified which leverages multiple knowledge bases in order to annotate literary content. However this multi-knowledge base approach is limited in the context of the challenges faced by this thesis due to the manner in which it uses these multiple sources.

In order to remedy problems with available knowledge that can inform an Entity Linking system, efforts are made to identify sources of knowledge which are not yet amenable to Entity Linking, but may prove to be helpful if they can be structured appropriately. Sources are identified both in the form of primary source and secondary source content. The secondary source content is structured into two ontologies which are subsequently linked back to DBpedia for the purposes both of leveraging its information in a multiple knowledge base Entity Linking solution and to facilitate integration between collections annotated with the new ontology, and those annotated with DBpedia. This linking process is performed automatically using a novel linking method.

Finally an approach to performing Entity Linking which combines multiple knowledge base sources is presented. A novel approach to constructing the knowledge base is presented. This approach facilitates both the use of and control over the multiple knowledge bases that inform the entity linker. It is demonstrated that this new system performs better than other tested systems when applied to various Entity Linking problems.

Contents

Glossary

Entity An entity is a distinct instance of some object in the world. It is an unambiguous, tangible "thing" which may be referred to, usually by name. In this thesis, entities are typically instances of people or locations.

Entity Linking The task of establishing a referent entity in a Knowledge Base for a given surface form. In this thesis it is synonymous with Named Entity Disambiguation. This task is described in detail in Chapter 2.

Entity Mention Sometimes used simply as "mention", this refers to a specific instance of a surface form. For example, the Catholic rebel Felim O'Neill may be referred to by the surface form "Phelim O Neal". Many instances of this specific surface form may be found in a document. Each instance is a mention.

Knowledge Base A resource which provides information about entities in a machine-readable format. The knowledge base informs the disambiguation system's choice of referent. It is usually semantic in nature, but this need not always be the case.

Knowledge Base Source A resource which can be used to construct a knowledge base. DBpedia is a common example.

Named Entity Classification The task of assigning a type to a surface form based on some set of predefined labels (usually Person (PER), Location (LOC), Organization (ORG) or Miscellaneous (MISC) e.g. Given the surface forms "Cú Chulainn", "Ulster", "Connacht" and "Táin Bó Cúailnge", a classification system should apply the labels PER, LOC, LOC and MISC respectively.

Named Entity Disambiguation The task of establishing a referent entity in a Knowledge Base for a given surface form. In this thesis it is synonymous with Entity Linking. This task is described in detail in Chapter 2.

Named Entity Recognition The task of identifying surface forms in free-text which refer to entities, e.g. given the string "Cú Chulainn defends Ulster from the army of Connacht in the Táin Bó Cúailnge", Named Entity Recognition should recognize and annotate the surface forms "Cú Chulainn", "Ulster", "Connacht" and "Táin Bó Cúailnge".

Referent The entity to which a mention is referring. The referent of a mention is usually identified by supplying a URI obtained from a referent knowledge base. This URI unambiguously identifies the referent.

Referent Knowledge Base The reference knowledge base is the KB from which a pool of candidate referents is retrieved during the EL process. When using multiple KBs as part of the linking process, the reference KB is some subset of the KBs being used.

Surface Form A string of characters which may be used to refer to an entity e.g. “Cú Chulainn”, “Ulster”, “Connacht”, “Táin Bó Cúailnge”.

Wikification Similar to Entity Linking except that the task is to link a given surface form to its corresponding Wikipedia article, rather than a referent in an arbitrary knowledge base. Wikification is essentially the genesis of modern Entity Linking.

Acronyms

BAT Benchmarking Entity-Annotation Systems.

BSD The Books of Survey and Distribution.

CH Cultural Heritage.

CHEL Cultural Heritage Entity Linker.

CR Coreference Resolution.

CRF Conditional Random Field.

DH Digital Humanities.

DIB The Dictionary of Irish Biography.

EE Emerging Entities.

EL Entity Linking.

HITS Hyperlink-Induced Topic Search.

IR Information Retrieval.

KB Knowledge Base.

LDA Latent Dirichlet Allocation.

NED Named Entity Disambiguation.

NER Named Entity Recognition.

NERC Named Entity Recognition and Classification.

NERD Named Entity Recognition and Disambiguation.

NLP Natural Language Processing.

ODNB The Oxford Dictionary of National Biography.

POS Part of Speech.

REDEN Referencement et Desambiguisation d'Entités Nommées.

RL Record Linking.

SVM Support Vector Machine.

VSM Vector Space Model.

WLM Wikipedia Link Based Measure.

WSD Word Sense Disambiguation.

XMI XML Metadata Interchange.

Chapter 1

Introduction

"Oh where is your inflammatory writ? Your text that would incite a light be lit?"

— Joanna Newsom, *Inflammatory Writ*

1.1 Motivation

Over time, many academic institutions have accumulated immense collections of cultural treasures, often dating back hundreds or even thousands of years. Unfortunately, it is often the case that accessing the contents of these collections is a challenging process for scholars. This may be because the artefacts are old and brittle. So much so that handling them may come at too great a risk to the integrity of the item. They may also be unreachable due to the immense volume and general disarray of the collections in question. Archivists and cataloguers have intricate systems for indexing and storing collections, but these systems have known limitations. It is not uncommon for a collection to be obtained by an institution and then simply forgotten. The artefacts become lost within a sea of treasures, awaiting rediscovery [34, 86].

By digitising these collections, distribution of the material is drastically simplified and the risk of damage to the medium is negated. The digitised documents can be shared and read by many people in parallel, theoretically supporting faster, more thorough research. Furthermore, by converting the documents into a format which can be parsed and understood by computers, machines may be employed to assist researchers in their investigations. A computer can construct and maintain an indexing system that is considerably more sophisticated than, for example, the well known Dewey Decimal system [22]. Given the right set of tools, it should be possible to immediately furnish scholars, both novice and expert, with artefacts that are of direct interest to them and their research questions.

The challenge is to index content in a manner that provides computers with an understanding of the complex structure and relationships inherent in these collections. Semantic Web principles and the promises that follow Semantic Web technologies [5] have taken root in the pursuit of solutions to this challenge.

Even as the idea of the Semantic Web was germinating, there were humanities scholars who realised that this technology would eventually become integral to humanities research [110]:

> In some form, the Semantic Web is our future, and it will require formal representations of the human record. Those representations – ontologies, schemas, knowledge representations, call them what you will – should be produced by people trained in the humanities. Producing them is a discipline that requires training in the humanities, but also in elements of mathematics, logic, engineering, and computer science.

Humanities scholars can gain direct control of how the computer manages and stores their data by defining formal vocabularies which describe an abstract model of cultural heritage items. As new items are added to the collection, scholars may annotate them with descriptive labels derived from the aforementioned vocabulary. In theory this increases the amount of support that a query interface can provide to a scholar and improves the manner in which digitised collections are organised and stored on disk.

However, as the rate at which collections can be digitised increases, and as the scale of digital cultural heritage collections grows, the manual effort of annotating and integrating new resources with existing collections becomes increasingly time consuming and expensive. There are many stages involved in digitisation and depending on the nature of the collection in question, different methods of automation may help to ease this process. It is here that we consider the possible applications of Entity Linking as a means to formally structure raw humanities text as it is being digitised.

Entity Linking (EL) is a problem in the domain of Natural Language Processing (NLP) which has seen much research over the last number of years. In brief, the goal of EL may be expressed as a mapping problem: **Given an input set of entity mentions, produce an output set of mappings which unambiguously establishes a single referent for each mention**. In other words, given an input text where references to entities have been extracted e.g. by manual annotation or by Named Entity Recognition (NER), the task of an EL service is to identify referents for each recognised entity.

To provide a concrete example, consider the following extract from *The Deposition of Henry Jones*[1,2], sourced from a collection known as the 1641 depositions which will be introduced in more detail in Section 1.4:

> I Henry Jones, doctor in divinity in obedience to his majesty's commission requiring an account of the losses of his loyal subjects wherein they suffered by the present Rebellion in Ireland, and requiring an account of what traitorous words, projects or actions were done,

[1] `http://1641.tcd.ie/deposition.php?depID=809001r001`

[2] For ease of reading the language of this extract has been normalised from Early Modern English, to adhere to modern spelling conventions.

> said, or plotted by the actors, or by the abettors in that rebellion: do make and give in this following report of the premises to the best of my knowledge upon oath...

Humans naturally excel at spotting mentions of entities in texts such as this. For them it is clear that this sample paragraph contains references to a number of entities: "Henry Jones", "his majesty's commission", "Ireland" and "Rebellion in Ireland". Humans also effortlessly classify them as a person, an organisation, a location and an event. Yet establishing exactly *who* "Henry Jones" is or *what* royal commission "his majesty's commission" refers to depends on the individual's personal knowledge and awareness of the context behind this sample paragraph.

For most people, "Henry Jones" might conjure to mind a character played by Séan Connery in the 1989 film "Indiana Jones and the Last Crusade" rather than a 17^{th} century Anglican bishop who investigated crimes allegedly perpetrated during the 1641 Irish Rebellion. Hence human associations between surface forms and referents are largely dependent on personal experience, acquired knowledge and cues offered by the source text. To use a heavy-handed and crude metaphor, a human's personal Knowledge Base (KB) affects the quality of the links produced by their internal EL system.

Similarly to humans, an EL system will begin the process of disambiguating a set of mentions by querying it's KB to retrieve a set of candidate referents. It will then use evidence derived from the candidates and the source text itself to assess the probability that any given candidate is the referent for the surface form. Assuming that Wikipedia is used as the KB (in which case the EL task is often called "Wikification"), a hypothetical EL system would indicate its chosen referents for each entity in this text by returning a link to a Wikipedia article that describes it. For the text above, these would be:

- **Henry Jones**: `http://wikipedia.org/wiki/Henry_Jones_(bishop)`
- **his majesty's commission**: `NIL`
- **Rebellion in Ireland**: `http://wikipedia.org/wiki/Irish_Rebellion_of_1641`
- **Ireland**: `http://wikipedia.org/wiki/Ireland`

Note the `NIL` annotation supplied for "his majesty's commission" in the above output. `NIL` is a special annotation which indicates that the system was unable to establish a suitable referent for the given mention. A `NIL` annotation may be supplied because there was not enough evidence in the source text to determine which of the candidate referents was the most likely choice or simply because no candidates could be found in the KB. In the case of this example, there is no Wikipedia article which describes the commission in which Henry Jones was involved (Commission for the Despoiled Subject), so the hypothetical EL system applies the `NIL` label.

This is an important limitation of EL systems which must be observed. Unlike humans who can gradually build up an awareness of entities as they encounter them in a source text, if an entity does not exist in an

EL system's KB, then it is simply disregarded as "unlinkable". The EL system does not learn or grow based on new information it encounters. Its view of the world is completely static and entirely dependent on the information it can garner from the KB.

The potential benefits of being able to automatically and accurately establish referents for mentions are numerous, but the most obvious advantage is content enrichment. An EL system provides a consistent identifier for each entity in the source text which can be referenced back to a central KB. While Semantic Web resources such as DBpedia [64] or YAGO [53, 108] are often the source of information for the KB, there is no reason why an ontology developed by expert historians could not be used.

This means that EL facilitates the integration of raw, flat textual content with the Semantic Web by applying semantic URIs to mentions. This enhances the content of an input text by linking it to new information available in other external sources. These enrichments may be used to inform users about the nature of a document's content or they may be used as inputs to some other automated process such as clustering, personalisation, information retrieval, question-answering etc. They can even help to build bridges across previously disparate collections as links between resources are found based on the mutual entities they contain.

It is important to highlight that these desirable outputs are for the most part predicated on the structure of the KB itself. Hence historians can control the ontological representation of collection entities, but the EL service facilitates annotation of newly digitised items according to this formal structure.

The problem with applying EL in the context of cultural heritage collections is that the corpora encountered generally require highly specialised knowledge in order to interpret their contents. As with the Henry Jones example above, an EL system that is informed by a generic KB such as DBpedia may require a push to use the more appropriate (but less popular) referent when tasked with linking the sample text. Aggravating the problem is the fact that the text being processed can be extremely noisy, either due to OCR errors or the archaic nature of the language used in the original artefact. Case in point, the sample text used in this section was normalised for ease of reading as interpreting the original 17^{th} century text might best be described as tedious (see Appendix A for sample documents).

Moreover, while these collections are vast from the perspective of a humanities scholar, they do not reach the scale that we have come to expect when working with big data. Although there is a temptation to apply deep learning solutions to problems, this may not be possible as the volume of training data available is simply insufficient for training deep learning models. Unfortunately we cannot simply deep learn the hard parts of our problems away when dealing with certain historical resources.

It is not outlandish to conjecture that the archives of the future will be almost entirely digital. Modern human history unfolds and is increasingly recorded in born-digital media. Enormous effort is expended to convert analogue collections into a digital format for the purposes of access and distribution, preservation, and enhanced investigation. Human history is not centralised and the artefacts which document the stories

of many significant historical events span multiple countries, or even continents. This is certainly true of Irish history [2, 17]. EL can provide a means of enriching these artefacts with supplementary information while also building bridges between disparate collections that would remain separate were they to be archived by other means.

1.2 Research Question, Goals and Objectives

The research question pursued in this thesis is:

> To what extent can entity linking be effectively performed on highly specialised text-based cultural heritage collections, such that an EL system can generate appropriate annotations for these textual corpora?

In the context of this research question "highly specialised" refers to the fact that domain specific knowledge is required on the part of an investigator in order to interpret the contents of a collection. This is analogous to the information available to an expert historian or literary scholar who would conduct low-level research into a given corpus. Standard EL KBs are unlikely to have sufficient coverage for this class of problem. Consequently this research will involve identifying and assessing the quality of available resources which might inform an EL system.

Careful consideration must also be given to the use of the word "appropriate". An appropriate annotation is more than simply a correct referent chosen from a pool of candidates. In the face of overwhelming ambiguity, sometimes the most appropriate action that an EL system can take is to abstain from providing an annotation. It is important that the outputs of an EL service are not misleading or heinously incorrect given the responsibility that has been assigned to the service.

On the nature of the textual corpora, there is an implicit assumption that the cultural heritage materials used in this work will be primary source collections. In particular, there is a focus on collections which use archaic, unnormalized or otherwise challenging spelling conventions. Working with such collections can be a taxing problem for off-the-shelf NLP tools and certainly creates challenges for existing EL systems.

The purpose of the research in this thesis is to identify and define resources, techniques and practices which can improve the application of EL techniques to cultural heritage collections which have undergone a digitisation process. This research is conducted with an eye to the humanities scholars who must interface with the outputs of such tools.

There has been much research into EL since its inception. Reducing the vast array of existing EL services to a core set of assumptions, features and solutions is the first goal of this work. Identifying which of these core approaches yields the best EL performance on our target content type is the second. Where

no single suitable approach is found, this thesis will investigate the creation of alternative or unifying methods that are more suitable to the content type targeted by this research.

There is a subtext to these goals, which is that the output of the EL process should be *a*) useful and trustworthy for expert historians *b*) informative for novice users, and *c*) well structured for the computer which must index and provide services built on the annotations. Although measuring impact on these three points is not within the remit of this research, it is a long-term objective which influences how the work is conducted. Accurate and effective EL is the goal of this thesis, but this objective exists in a much greater sphere of research which aims to facilitate better organisation and accessibility with respect to digital cultural heritage collections.

1.3 Research Contribution

This research provides a set of techniques for applying EL to noisy text-based cultural heritage collections. This includes resolving issues with poor coverage in the knowledge base, providing considerations for abstaining from annotating where appropriate and compensating for disparities in surface forms between the KB and entity mentions. These approaches are evaluated quantitatively using standard benchmarks that are widely adopted by the EL community. While the result is not a "perfect" EL solution, it demonstrates a significant improvement over the performance of other off-the-shelf tools that were investigated.

A contribution of this work is a novel approach to constructing a KB for EL in a manner that facilitates linking across multiple KBs while being able to control which KB sources should be used. This has resulted in the creation of the Cultural Heritage Entity Linker (CHEL). CHEL uses graph-based heuristics and record linking similarity measures to perform disambiguation on noisy text-based cultural heritage collections. Demonstrable improvements in performance over the state-of-the-art [40, 112, 9] are achieved by harvesting and resolving information from multiple sources in order to tackle challenging Cultural Heritage (CH) collections. The performance of this system is tested using three different corpora – a subset of Europeana content, 17^{th} century legal documentation and a collection of French literary criticisms.

CHEL is implemented as an e-service for the FREME-NER project meaning that it can be integrated with existing deployments of the FREME entity linking service online. CHEL is fully open sourced, highly configurable and may be obtained from Github[3].

An observable problem when linking Irish cultural heritage collections is the lack of suitable KBs which provide adequate coverage for the entities encountered. A second contribution of this research is the creation of two foundational ontologies for linking notable Irish people who are historically significant, but not documented in more general ontologies such as DBpedia, Wikidata, Freebase, and YAGO2 [64, 8, 53, 114].

[3] https://github.com/munnellg/CHEL

Crucially, there are no published ontologies for describing Irish people through history that we are aware of. Christopher Yocum's work on the Irish Genealogies Ontology[4] (discussed in Section 2.6.5) is an important contribution that warrents consideration, but this source only covers medieval Ireland, while the two sources produced by this thesis span approximately 1500 years.

During the construction of the specialised KB, a method of automatically integrating newly generated, specialised knowledge bases with more popular, generic alternatives was designed and evaluated. This constitutes another contribution of this research.

Finally, a problem with investigating EL in cultural heritage is the lack of available datasets which may be used for evaluation. Another contribution of this thesis is the creation of a new gold standard EL dataset. This new collection presents some of the unique challenges faced when performing EL on Irish historical resources.

A list of publications based on this research are given below:

Long Papers

Munnelly, Gary; and Lawless, Séamus. "Investigating Entity Linking in Early English Legal Documents". In the Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018. June 3rd–6th 2018. Fort Worth, TX, USA.

Munnelly, Gary; and Lawless, Séamus. "Constructing a Knowledge Base for Entity Linking on Irish Cultural Heritage Collections". In the Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018. September 11th–13th 2018. Vienna, Autria.

Workshop Papers

Munnelly, Gary; Pandit, Harshvardhan; and Lawless, Séamus. "Exploring Linked Data For The Automatic Enrichment of Historical Archives". In the Proceedings of the 3rd International Workshop on Semantic Web for Cultural heritage, SW4CH, held in conjunction with 15th Extended Semantic Web Conference, ESWC 2018. 3rd–7th June 2018. Heraklion, Crete, Greece.

Munnelly, Gary, Caputo, Annalina, Lawless Séamus. "Linking Historical Sources to Established Knowledge Bases in Order to Inform Entity Linkers in Cultural Heritage." In the Proceedings of the 1st Workshop on Computational Methods in the Humanities, COMHUM 2018. June 4–5, 2018. Lausanne, Switzerland.

Collaborative Research

In addition to the core focus of this thesis, research also yielded an interesting collaborative project into the use of EL as a means of tracking the context in which entities are mention changes over time in news

[4] https://github.com/cyocum/irish-gen

collections. Detail on this collaboration is given in Section 6.3.2. This work with Caputo resulted in the following publication:

Caputo, Annalina; Munnelly, Gary; and Lawless, Séamus. “Temporal Entity Random Indexing”. In the Book of Abstracts of the 28th Digital Humanities Conference, DH 2018. June 26th–29th 2018. Mexico City, Mexico.

1.4 Research Approach

The main focus of the research question is on investigating the creation and improvement of EL methods such that they can be used to reliably annotate extremely challenging CH material. EL lends itself well to a quantitative method of evaluation as performance can be computed by using metrics such as Precision, Recall and F1 Measure (a more detailed discussion of this process is provided in Chapter 2).

The challenges faced within cultural heritage are relatively unique due to the immense variation in the quality and language of the textual content. This thesis initially uses a corpus known as the 1641 depositions[5] for evaluation. The depositions are a textbook example of the kind of challenges EL faces in cultural heritage.

The 1641 depositions are a collection of 8,000 depositions or witness statements, examinations and associated materials, amounting to 19,010 pages and bound in 31 volumes. They are written in archaic English making them extremely noisy. The use of language is inconsistent – the entity “Devil” for example has multiple spelling variations, including “Diuill”, “Divil”, and no instances of the modern spelling – and ancient naming conventions make resolving entities to their modern equivalents challenging.

A gold standard annotated subset of the 1641 depositions has been produced and is used to perform a baseline evaluation of existing EL systems. The goal of this evaluation is to identify which existing EL services exhibit the most robust performance given the challenging nature of the corpus.

Once the most effective approach to EL has been identified, a deep analysis of its inner machinations is conducted to identify the design decisions which make it efficacious for the chosen corpus. The limiting factor of the KB is also considered and an investigation is performed to identify alternative knowledge resources to inform the EL system. Suitability of a resource is quantified based on the coverage it provides for the target corpus.

A new EL system is developed based on an analysis of effective existing solutions and the discovered limitations of those solutions with respect to the CH content that is of interest in this research. This new system is applied to the same gold standard corpus under different configurations in order to determine if the approaches proposed by this research are effective.

The transferability of the new system is also considered by evaluating it with respect to two other cultural

[5] http://1641.tcd.ie/

heritage collections, namely a collection of French literary texts and a collection of content descriptions harvested from Europeana.

1.5 Thesis Outline

What follows is a brief outline of the structure of this thesis and the information that can be found in each chapter. There is a general continuity which follows Chapters 3, 4 and 5 meaning it is advised that they should be read in sequence.

Chapter 2 presents the state-of-the-art on two fronts with respect to EL. First, traditional EL is discussed along with several perspectives and solutions to the EL problem when it is tacked with the objective of annotating contemporary textual content. The latter half of this chapter specifically discusses efforts to apply EL technology to the challenge of annotating CH content. Some notable contributions in this field are discussed in detail.

Chapter 3 presents a comparative evaluation of EL systems using a corpus of 17^{th} century Irish depositions. It is shown that state of the art EL systems struggle to annotate the depositions. Yet some desirable properties of an EL system are identified and studied in detail. This chapter also serves to highlight the challenge faced when identifying KB sources that can adequately annotate a corpus such as that used for the experiment.

Chapter 4 tackles the problem with entity coverage in the KB by identifying sources of information that are previously untapped and are highly relevant to research in Irish CH. Both primary and secondary source material is investigated and two new ontologies are created from secondary source material. Additionally a linking method for determining equivalent entities between KB sources is presented and used to create links between the new ontologies and DBpedia.

Chapter 5 presents a novel approach to performing EL using multiple KB sources simultaneously. The linking method is presented and then rigorously evaluated against three different content types. Conclusions about the effectiveness of the method are drawn based on the results of these experiments.

Finally Chapter 6 presents a summary of the findings of this thesis, restates the goals and contributions, and highlights how these were met through research conducted in previous chapters. Several interesting avenues for future research are suggested.

Chapter 2

State of the Art Entity Linking

"Wait, wait, let me explain something to you. I am not Mr. Lebowski. You're Mr. Lebowski. I'm The Dude."

— The Dude, *The Big Lebowski*

2.1 Overview

This chapter will present a discussion of the current state of the art in the field of EL. For this thesis, the discussion must be carried out in two parts: EL as a field of study, and EL as it has been applied to problems in CH.

The first half of this chapter provides an introduction to EL. The components of an EL system are described as well as the means by which they may be implemented. Methods of evaluating the performance of an EL system are also described.

The latter half of this chapter examines the application of EL to problems in CH. These approaches are, for the most part, quite simple. However, they illustrate the difference that exists between how EL systems are normally researched and implemented for general use, versus how they are typically deployed in a highly specialised setting.

2.2 An Introduction to Entity Linking

From the perspective of a computer, human language is an inconsistent and often ambiguous means of communicating information. Sometimes it is remarkable to consider the ease with which the human brain processes and creates associations between words and tangible entities. Even in the absence of surrounding information, it is usually possible for a human to assign some mental referent to a string of text based on prior knowledge, personal bias or some perceived relationship between multiple instances of surface forms.

Consider, without any surrounding context, the surface forms “Kepler”, “Mars”, and “Tycho Brahe”. Given no other context than this, what might these strings be referring to? “Kepler” could be a reference to a spacecraft, a star or a 17^{th} century mathematician . “Mars” might refer to a celestial body, a popular confectionery or a Roman god of war. “Tycho Brahe” could be a 16^{th} century astronomer or a Martian crater.

If a referent for each of the three surface forms must be chosen from the options given above and only one configuration of those chosen referents is correct, how should the referents be chosen? If it is assumed that the surface forms are related by some underlying theme, then it is possible to make an educated guess that Tycho Brahe refers to a crater and Mars refers to a planet, given that the crater Tycho Brahe is located on the planet Mars. However, this does nothing to help identify Kepler. It could be argued that the choice of the mathematician, astronomer and planet are sensible as Kepler’s work in relation to Mars is based on the work of Tycho Brahe. So perhaps this choice makes sense.

This example is clearly contrived as ordinarily surface forms are observed within some sort of context enabling the identification of the subject of a reference. In the case of this example, the surface forms have been extracted from a passage in Richard Feynman’s “The Character of Physical Law”, with the passage being shown in Figure 1. Given the surrounding context it is possible to identify that the correct referents are indeed the mathematician, astronomer and planet.

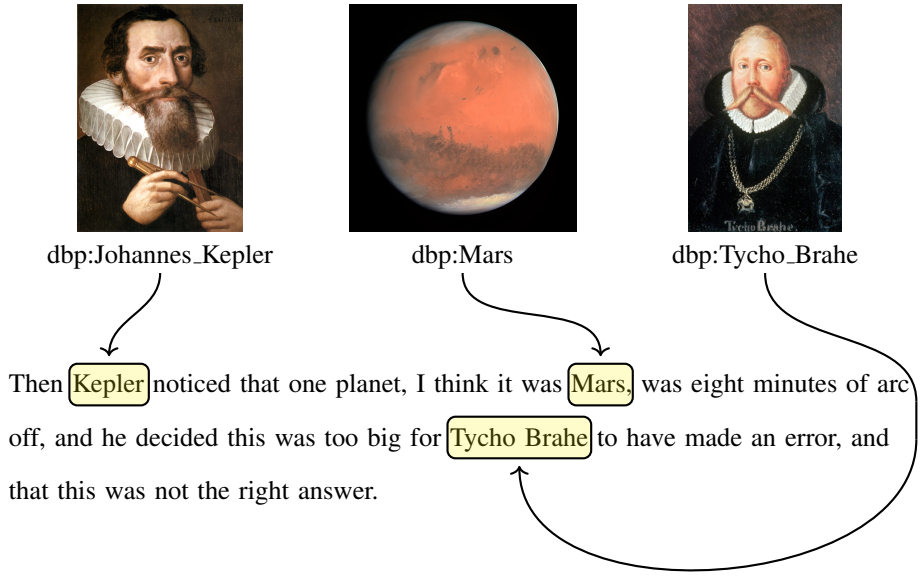


Figure 1: Choosing referents for mentions identified in context.

The Entity Linking (EL) problem in NLP is effectively a document annotation problem, whereby a computer is challenged with performing precisely this task of assigning appropriate referents to an input series

of surface forms based on some set of features that are chosen by the programmer.

A user submits a document containing ambiguous mentions of entities. The EL system queries a source of information known as the KB, which contains information about all entities that are known to the EL system. From the KB a pool of candidate referents is generated for each ambiguous mention in the input text. A referent selection process proceeds to choose an appropriate referent for each mention. The identity of the referent is established by supplying a semantic URI obtained from the index of entities in the KB. Where no suitable referent is found, the EL system applies a generic NIL annotation (or simply supplies no annotation) in order to denote that no suitable referent could be found. The annotated mentions and their corresponding URIs are returned to the user.

An immediate benefit of performing EL on a corpus is that it imposes structure on the contents of a collection, facilitating the categorisation, organisation and interlinking of documents using semantic URIs. The collection is automatically uplifted from a raw, text-based representation to an intelligent, semantic format.

Secondary benefits may be derived from the results of EL through the application of a range of technologies including semantic search [51, 84], or other novel forms of analysing content [11] .

EL has seen much research over the last number of years, and there are a diverse range of creative solutions. Some of these solutions are extremely simple, while others employ sophisticated methods of measuring the strength of relationships between candidates, or the relevance of the contexts in which candidates are expected to be found. Some of these approaches will be explored throughout this chapter.

2.3 Why Semantically Annotated Data Is Better

Sections 1.1 and 2.1 have both stated that the annotation of cultural heritage material with semantic URIs is an advantageous decision, but the case has not yet been made as to *why* this is a good decision, or what the application of "Semantic Web Principles" yields for the curator of digital cultural heritage content. This section aims to elaborate on some of these statements.

Consider the case of data as it is stored in a database (MySQL, MongoDB etc) or an Excel spreadsheet. It is easy to immediately identify a number of problems with these traditional approaches to curating information.

First, it is the task of the human to understand what each entry in a given database represents and the subsequent implications for the stored data. For example, a spreadsheet could be used to represent people mentioned in a collection of documents, with each row representing an individual person. If the researcher has chosen to capture information such as date of birth, name and occupation, then all curated information about a person can be found by reading the appropriate row in the spreadsheet.

However, as humans, we implicitly know that each person may also have familial connections with other

entries in the spreadsheet. We know that every person must have a biological mother and father. We know that they most likely had a place of residence, with that place of residence itself being a complex entity with multiple interesting properties and qualities. But how can the computer understand that? How can the computer distinguish between a storage medium whose records describe people, and one whose records describe locations? How can the computer be aware of surrounding contextual information that is absent from this particular representation of entities?

This problem extends beyond the limitations of a computer's comprehension. Suppose an individual constructs a database of information before subsequently leaving an organisation. The database is inherited by a second individual. Can the new individual understand the original curator's approach to storing information? Is a database schema expressive enough to convey to this new individual what the various relationships between tables represent? Clearly not as a database's relationships between tables are limited to simple foreign keys. We must hope that the original curator was descriptive enough in the naming of those keys so that their semantics are conveyed to the human. Assuming best practice was used, the semantics of the relationship are still lost on the computer itself.

Yet another observation is, how can we reference entities from the multitude of disparate spreadsheets, databases and other storage mechanisms that host our curated information about the past? Suppose we were to curate a collection of 17^{th} century maps and we have taken the time to extract all locations and store them in a database. We would now like to connect this new database to an older database of depositions which describe crimes committed in the same time period. How can our new database reliably reference the old one? Do we need to fully ingest the older database into the new one in order to ensure that semantic constraints such as foreign keys are appropriately respected?

These challenges represent some of the problems that can be mitigated through the appropriate application of semantic web principles. Describing this solution is a discussion comprised of two parts – the ontology and the knowledge graph.

An ontology is an abstract description of data. In the simplest terms it could be said to describe a taxonomoy of entities (although the reality is deeper than this), denoting that a person is a sub-class of a biological object, for example. The ontology describes properties that an entity can have, such as date of birth, parents etc. It also imposes clear constraints on the values that these properties can have (the domain and the range). The ontology is analogous to the traditional database schema, but it is both more expressive in that it can be used to describe the complex semantic implications of relationships between entities, and it is more flexible, as the extension of the ontology to meet new challenges can be achieved by altering the ontology without the need to alter the data itself.

The knowledge graph represents a populated datastore which uses the ontology to describe its contents. Information is expressed as triples in the form (Subject, Predicate, Object), thus forming a graph of information. The knowledge graph references the ontology in order to describe its data, but it does not need to contain the ontology. Again, this allows the ontology and the knowledge graph to evolve

separately. The knowledge graph is the component that we are most concerned with in this thesis. Given a populated knowledge graph, it should be possible to link new information found in digitized documents to the appropriate instances of entities in the graph.

According to best practices, the contents of both the ontology and the knowledge graph should individually be given a URI which is resolvable over the web. This is higly advantageous. When referencing entities, we can be completely unambiguous about the entity to which we are referring by supplying its URI which should be globally unique. An individual who wishes to learn more information about the entity can do so by browsing to their URI over HTTP.

When modelling data with the ontology, statements made using the ontology are expressed again through the use of their URIs. If the ontology evolves or changes, these changes can be immediately reflected in the data because the schema and data are separate. As the schema evolves, the data can keep up, because it only has to reference the parts of the ontology that it is using.

In the event that we wish to introduce a huge shift in the nature of the information we are curating, e.g. extending our knowledge graph of people to incorporate official titles, offices and occupations, this can be achieved either by extending the original ontology *or* simply introducing a reference to a second ontology which models this information. Indeed, growing our knowledge graph to meet new challenges is simply a task of finding an ontology that is expressive enough to meet our new demands, and referencing it where appropriate in the knowledge graph.

The ontology also facilitates inference over the contents of the knowledge graph. Previously it was mentioned that a human reading a spreadsheet of information about people knows that there is a vast amount of surrounding information that the spreadsheet does not capture. While our knowledge graph may capture shallow information such as name and date of birth for a person, the ontology tells the computer that instances of a person entity will also have biological parents. Furthermore these biological parents will, themselves, be people with dates of birth etc. Should these individuals be identified, they may be referenced by linking their URI to the URI of their child via the appropriate property as described by the ontology.

Of course, there are disadvantages to the use of ontologies and knowledge graphs too, with the greatest being that they can be more challenging for users to understand, and require greater effort to populate and maintain than traditional databases. As ontologies try to be as broad as possible in the range of information they can describe, this can result in vocabularies that are extremely verbose when attempting to describe simple concepts. This is why automatic means of populating and maintaining these resources can be so valuable from the perspective of the user.

Because the ontologies must be hosted online, a problem also arises when an institution can no longer afford to host their ontology or knowledge graph. What happens when a key ontology disappears from the web? Answers to this problem form the basis of ongoing work within the semantic web community.

2.4 What is Entity Linking and What is Not

There is some inconsistency in the literature with regards to what exactly is within the remit of EL. For example, some argue that EL is a synonym for Named Entity Disambiguation (NED), simply involving the identification of a referent for an input set of surface forms [44, 9, 91] while others claim that it is in fact synonymous with Named Entity Recognition and Disambiguation (NERD), involving both spotting surface forms in free text, and the identification of referents for those spotted entities [4, 116]. It has also been said that a difference between NED and EL is that an EL system does not provide NIL annotations while a NED system does [4].

This thesis follows the former definition, i.e. EL is synonymous with NED. Let it be stated that for the purposes of this thesis, Entity Linking is defined as the following task:

> Given an *input set of entity mentions*, the challenge in EL is to identify a corresponding mapping of *mentions to referents* where referents are obtained from a *semantic knowledge base*. Where no suitable referent exists in the KB, a *NIL annotation* should be applied to the corresponding mention.

Given this concrete definition, it is worth considering related fields that are sometimes confused with EL or can provide insight that complements EL. These fields are discussed below.

2.4.1 Named Entity Recognition and Classification

Named Entity Recognition (NER) and Named Entity Recognition and Classification (NERC) are two different problems which are often performed together, so much so that NER is sometimes implicitly assumed to mean NERC.

NER is the task of identifying which tokens in a stream denote references to an entity. The precise types of these entities is not considered. Trivially the positions of entities can be identified using Begin-Inside-Outside (BIO) labels with all tokens that are not entities being marked O (outside), tokens which denote the start of entities being marked B (begin) and tokens which continue a reference to an entity after an initial token that was marked B are given the label I (inside).

Solutions to NER typically involve training a classifier over sequences of tokens. These tokens may be annotated with the aforementioned BIO labels. The classifier must learn to classify tokens as being indicative of an entity mention or not. In Stanford NLP for example a Conditional Random Field (CRF) is used [33]. The Illinois NER tagger uses a neural network with beam search [92].

Showing that there is more than one solution to the NER problem, however, NLTK has functionality to define a regular expression based NER tagger where certain sequences of Part of Speech (POS) tags are deemed to denote entities [6].

NERC involves both NER and then the subsequent typing of tokens that have been marked as entities. Typically a NERC system applies one of four labels to a recognised entity: Person, Place, Organisation or Miscellaneous. Collectively these are known as ENAMEX types [42]. While it can usually be assumed that a NERC system will use ENAMEX type labels, research has been conducted into the application of more fine-grained entity types [54] which are supported by some NLP toolkits e.g. SpaCy[1].

NER and NERC are generally considered to be fundamental NLP tools and every NLP software package that was experimented with during this PhD provided NERC in some format.

As was previously mentioned, some researchers argue that NER is a task which is included in the EL process. However, this thesis sides with the argument that they are to be considered separately due to the fact that they can easily be decoupled and solved in isolation. NERC is useful for EL, but solving NERC problems does not directly help to solve EL problems.

2.4.2 Coreference Resolution

Coreference Resolution (CR), like EL, is a problem which involves resolving ambiguous mentions of entities in order to identify a single referent. Unlike EL, there is no linking to an external knowledge source. The entities are derived from the collection in question and the goal is to determine which mentions can be clustered as references to a specific individual.

It is important to be clear on the difference between EL and CR. Although, at a high level, their respective outcomes may seem the same (i.e. the resolution of entity mentions using some consistent referent) ultimately the manner in which they are implemented and the nature of the information they output is different. Yet they can easily be confused with one another by researchers who are unfamiliar with the domain. Hence the following description of CR is provided as a means of emphaising how it is distinct from EL.

For example, given the following sentences, the challenge in CR is to identify that the “he” in the second sentence is a reference to “John Romero” in the first sentence:

> John Romero is best known as a co-founder of id Software. He currently resides in Galway, Ireland.

According to Piskorski et al. [88] CR can be broken down into four types of resolution:

- Named: the resolution of proper nouns e.g. “John Romero”
- Pronominal: the resolution of pronouns e.g. “his”, “her”
- Nominal: the resolution of phrases e.g. “the company”

[1] `https://spacy.io`

- Implicit: the resolution of zero-anaphora, where the anaphora is indicated by a break or gap in a statement e.g. “Joe’s new car looks amazing, all shiny and chrome”. The latter clause refers to Joe’s car without the use of an explicit anaphora.

CR can be useful as a means of finding repeated references to a specific entity throughout a collection or even as a means of identifying which subsections of a larger document are relevant to discussion about a single entity. The benefit for a single collection is somewhat similar to that obtained with EL. Namely it is possible to group and categorise documents based on the nature of the entities they contain and this in turn can be used to provide more meaningful search etc. However the collection sits in isolation and cannot communicate with other external collections unless some form of CR is performed between their contents. Due to the semantic URIs used in EL, it becomes immediately possible to connect a newly annotated collection with other external collections that use the same vocabulary.

2.4.3 Record Linking

Record Linking (RL) is a problem which is sometimes confused with EL, although it is an entirely different challenge altogether. RL is the problem of creating mappings between records which describe the same entity in two separate databases [118, 100, 27].

For example, imagine there exist two separate databases. One contains census data about people who lived in Ireland in the year 1670. The other contains records of loans owed to bankers in Ireland. A researcher wishes to identify how many of the individuals recorded in the census were suffering from financial distress due to banking loans. RL solutions would attempt to map records between the two databases where a given banking transaction can be assigned to a person who is listed in the census.

One of the more well known models for performing RL is the Fellegi–Sunter method [25]. This approach identifies attributes that are present in both sets of records (e.g. in the previous example this may be a name and address field) and computes a log-probability score based on the similarity of these attributes. The definition of similarity depends on the nature of the fields being considered, but the application of simple measures such as Levenshtein distance [65], Dice Coefficient [23], Jaccard Index [57], Jaro-Winkler distance [117] etc. are common.

RL tools can be extremely useful for CH researchers as the digitisation and analysis of CH collections can result in the construction of multiple databases of records which contain some degree of mutual information, but do not contain references to each other. Where the decision is made to resolve these databases, the process can be highly manual and time consuming. RL can help to expedite this process.

2.4.4 Word Sense Disambiguation

A closely related field to EL is Word Sense Disambiguation (WSD). Unlike EL which typically focuses on named entities, WSD tackles the problem of identifying the sense of a word as it appears in context.

For example, given the input "soldier", does the word in a given context refer to a member of the army, or a strip of toast that is served with eggs.

The challenges in WSD are quite similar to those found in EL and some of the solutions between the two problems mirror each other quite closely. In fact, during the early days of EL research it was noted by Hachey et al [44] that WSD graph-based techniques could be applied to the EL problem. After generating candidate entities, Hachey selected any surface forms which had only a single candidate disambiguation, i.e. mentions that were already unambiguous. A graph was seeded with the candidates of these unambiguous surface forms and a new subgraph of of the KB (Wikipedia) was generated based on the links these pages contained.

Perhaps the greatest difference between EL and WSD are their respective assumptions with regards to the KB. WSD generally assumes a closed world, meaning a language is comprised of a series of words and all of these words can be listed and accounted for. This has led to incredible displays of human effort which attempt to document languages and the relationships between words [28, 72].

The open world assumption in EL states that no matter how thoroughly documented entities may be, there will always be some that are not present in the KB. This is important, because it means that at some point, inevitably, an EL system must be able to hold its hands up and declare that it does not know of any suitable referent for an input entity. Generally this is done by applying a generic NIL annotation to mentions that are believed to refer to out-of-KB entities.

EL and WSD are not at all mutually exclusive. Babelfy [78, 77] for example is a combined EL/WSD system which performs both tasks in unison. The two challenges complement each other well and while they are considered separately for the purposes of this thesis, there is value to be found in reading literature that tackles the WSD problem.

2.4.5 Wikification

Wikification is very similar to EL in that it involves the identification of referents for named entities in text using an external KB as a source of identifiers. The two challenges differ based on the nature of their respective KBs.

Wikification is essentially the challenge of linking mentions of entities in free text to their corresponding articles in Wikipedia [32, 87]. Systems which perform linking exclusively with respect to Wikipedia have the advantage of making certain assumptions about the nature of the KB. They can assume the existence of hyperlinks between articles, long form descriptions of entities, user statistics which indicate the popularity of an entity and more. Wikifiers can also make use of Wikipedia categories, redirects, and disambiguation pages. These are powerful features, but they tightly bind the linking system to Wikipedia.

The roots of EL can, in fact, be traced back to the work of Bunescu and Paşca [10], who realised the immense potential of Wikipedia as an aide in solving the problem of linking ambiguous references in text

to their respective referents. Their method treated individual Wikipedia pages as entities and attempted to link mentions in text back to the Wikipedia page of the entity they referred to. Referents were identified using the URL of the corresponding article.

By leveraging Wikipedia's disambiguation and redirect pages, Bunescu and Paşca were able to build a disambiguation dictionary which mapped surface forms to candidate entities. Using a combination of Wikipedia categories and context windows around mentions as features they established a measure of similarity which allowed them to distinguish between different entities which had the same surface form based on how they appeared in text.

Two methods of computing the similarity between entity mentions and candidate disambiguations were tested – Cosine similarity between the text of an entity's Wikipedia article and a surface form's context (as determined by a context window around the surface form in the document), and a taxonomy based Support Vector Machine (SVM) [60] trained on information mined from Wikipedia hyperlinks.

Later Cucerzan expanded on their work and introduced the idea of coherence between entities, reasoning that entities which are mentioned in the same document are likely related by some common subject matter [18]. In order to measure the agreement between entities Cucerzan created sets of vectors which described entities in terms of their context and the categories to which they belonged. With these vectors established, document vectors could be derived for input texts. A document vector was comprised of the contexts it contained and the categories assigned to the candidate disambiguations of each surface form. He then treated the problem as an optimisation problem and tried to find solutions which maximised both the agreement between an entity context vector and the document context vector as well as the agreement between the entity category vectors for each chosen disambiguation. Agreement was computed as the dot product between vectors.

Milne and Witten [74] exploited the link structure of Wikipedia to compute a measure of semantic relatedness between articles [119]. They trained a C4.5 decision tree [89] using a number of features derived from these links, which included a combination of the popularity of an entity as a disambiguation candidate, the confidence of the system in its disambiguation choice and the homogeneity of contexts suggested by disambiguation candidates. This contribution was particularly interesting because it moved away from previous attempts to link entities based on their individual textual similarity and moved towards an approach that focused on the global coherence of candidate entity referents (see Section 2.8 for a descriptions of local and global coherence).

Recent work by Nanni et al. [81] has taken the Wikification task further, attempting to link mentions to specific "aspects" of themselves in a Wikipedia article. For example, given the mention "Hillary Rodham Clinton", Nanni returns a link, not only to the corresponding Wikipedia page[2], but to a specific subsection of the article which identifies which temporal aspect of Hillary Clinton is being discussed.

[2] `https://en.wikipedia.org/wiki/Hillary_Clinton`

Wikification is not limited to the annotation of entities. The Wikify! tool developed by Mihalcea et al. [70] is a WSD tool that uses Wikipedia articles to identify the sense of a word being annotated. As with named entities, the service attempts to identify a Wikipedia article which can best explain the sense of a word as it appears in context.

Wikification is a separate task to EL, but there is ample overlap between approaches to Wikification and solutions for EL. Indeed, the origins of EL very much to lie in Wikification [10, 74] and modern approaches to EL are heavily inspired by the original research conducted using Wikipedia as a KB. None-the-less the fields have diverged on the basis of what is respectively used as the KB and it was felt when writing this thesis that it was important to distinguish Wikification from EL.

2.5 Architecture of an Entity Linking System

A typical EL system can be broken down into three core components which are discussed in detail below in Sections 2.6, 2.7 and 2.8. These components are the Knowledge Base, the candidate selection process, and the referent selection process. The relationship between these components is shown in Figure 2.

In this design, a user submits a corpus of documents to be annotated by an EL system. The corpus may be comprised of multiple documents and may be supplied with accompanying configuration data which is provided by the user. The nature of this configuration depends on what options the EL service exposes to the user. It may contain settings for threshold confidence scores that determine whether a mention is assigned a referent or NIL. It may choose the KB from which URIs are to be sourced. It may determine the referent selection method that the EL service should use. It is assumed in this diagram that the entities in the text of the corpus have already been annotated.

The mentions in the corpus text are passed to the candidate selection process which iteratively executes them as queries against the KB. Each mention is considered in turn and the KB returns a set of candidates based on some selection method. In many EL implementations, the KB is a Lucene search index and the selection method is the default relevance scoring metric used by the search engine library. Relevance is computed between the mention and surface forms in the KB.

Note that the KB is constructed based on some third-party source of information. In the case of this diagram, the source is DBpedia but any semantic ontology which contains relevant information about entities could be substituted here.

The candidate selection process aggregates the candidates that are identified in the KB. It may filter some candidates, as discussed in Section 2.7. The pool of candidates is then returned to the referent selection process which assesses the quality of candidates as referents for each mention. Either a URI or a NIL label is applied to each mention as appropriate and the results are returned to the user.

A discussion of EL which covers these three major components is reasonably complete. However, the

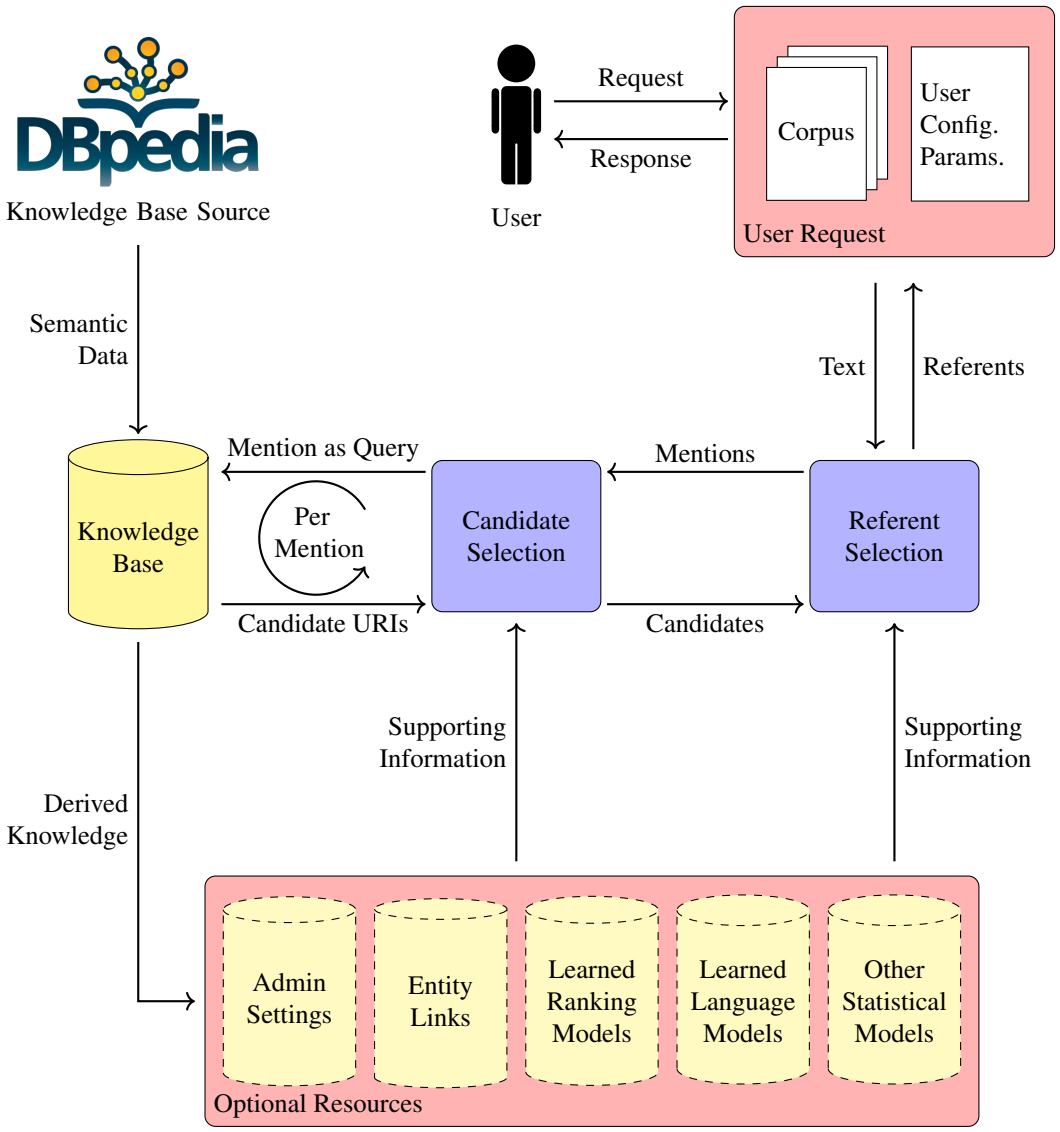


Figure 2: Example EL system architecture.

diagram in Figure 2 attempts to capture the fact that there is a great deal of surrounding information which supports (or rather is built into) these components. The referent selection process, for example, must be present in all EL systems. But how does it arrive at a choice of referent for a given mention? Different EL tools will avail of different supporting information in order to make this decision. This is captured by the "optional resources" component of the diagram.

Optional resources is a placeholder for whatever pre-processed information is available to the referent selection process or the candidate selection process whilst assigning referents to mentions in a corpus. It may, for example, contain probability priors which represent the likelihood that a given mention refers to an entity, based on how often a surface form that matches the mention was used as anchor text that linked to the corresponding entity (represented as "Other Statistical Models" in Figure 2). This information is actually used during the candidate selection process in some EL systems [68].

During the design of an EL system, perhaps a language model is constructed which can be used to compare the context of an input document to known contexts of entities in the KB. The candidate referent with the greatest contextual similarity to the input document is the most likely referent.

The diagram includes "admin settings" under this umbrella of optional resources. These are specific, hard coded decisions made by the maintainer of the EL system. These settings may limit the EL system to only linking entities that belong to a specific type (people, places etc.) or may determine a geographic boundary from which all location URIs must be sourced. They are the conscious decisions made by a developer based on the intended use of the EL system.

Relationships between entities are a common source of information used to select referents. This knowledge is often derived from the KB itself and is represented by the "entity links" module in the architecture diagram.

"Learned ranking models" refers to the fact that some EL systems treat the linking process as an Information Retrieval (IR) problem. The organisation of an EL system is not entirely dissimilar to that of a simple text-based search engine. Multiple queries (entity mentions) are executed simultaneously with respect to a search index and the objective is to rank the correct referent for each mention as the first result in the individual result sets. The result set for one query can (but does not necessarily) influence the result set for a separate query. When thought of in this way, it is easy to see how standard IR methods may be extended or employed to construct an EL service. For example, some EL systems use a learning to rank approach to identify referents [123].

Indeed, in FREME-NER [99], the search engine Solr effectively *is* is the EL system with each mention being executed as a query against the Solr API and the top ranked result being returned as the referent.

The purpose of the optional resources module is not to be an all-encompassing list of every kind of resource that supports an EL system. Rather, its function is to highlight some of the complexity that sits behind the seemly innocuous referent selection and candidate selection processes. Essentially, it is far

too easy to simply treat EL systems as a black-box and not worry about how the individual components actually function or what information drives their decision process. This is a problem and the objective of this diagram is to highlight that EL systems are complex, with many underlying components whose nature, or indeed very presence may vary from implementation to implementation.

2.6 The Knowledge Base

The Knowledge Base (KB) represents all the information that is available to an EL system when identifying referents for an input text. It documents instances of entities, names by which the entities may be identified, relationships between entities and more.

In this thesis there is a distinction made between a KB and a KB source. The KB is an indexed, pre-processed repository of information that is directly queried by the candidate selection process during EL. A KB source is a source of information which was used to construct the KB. The KB may be a filtered subset of information in the KB source or it may be the result of applying some transformations to the data that the KB source contains. For example, DBpedia is described in this thesis as a KB source. A search index which indexes DBpedia for the purposes of EL is a KB. Admittedly this is not common vernacular, but it was felt that a clear distinction needed to be drawn between sources of information, and the manner in which those sources were ultimately used.

While the KB source is usually semantic in nature, this does not necessarily need to be the case, given that early EL systems used information mined directly from Wikipedia to inform the linking process. Arguably the purpose of an EL system is to semantically annotate raw text using meaningful identifiers, and for all intents and purposes it can be assumed in modern applications that the KB source is semantic. However, when considering how specific a CH collection can be, it is worth considering the possibility that a semantic KB source does not exist and linking may need to take place with respect to, perhaps, a relational database until such time as a semantic KB source can be constructed and used.

Inevitably there will be gaps in the information provided by a KB source, as there are in all sources of information. There is no single ontology of all things. Different KB sources will have different target applications and, by extension, different limitations with respect to the information they provide. This can be desirable in certain contexts. It is good to have a KB that is tailored to a specific problem space, rather than requiring that a referent selection process trawl through masses of irrelevant information. A tailored KB reduces the possibility of inaccuracies in the output annotations and can cut down on the execution time of the EL system as it has less ambiguity to resolve.

A selection of KB sources relevant to this research have been presented below.

2.6.1 Knowledge Base Sources Derived from Wikipedia

The DBpedia ontology [64] is a repository of information about entities that has been constructed by automatically extracting structured information from Wikipedia articles (e.g. by examining the content of infoboxes) using a comprehensive extraction framework[3]. Detailed information about this extraction process is provided in [64].

DBpedia contains information about entities in 111 languages and makes its data available as a series of datasets that focus on a particular type of information extracted from articles e.g. anchor text, geographic coordinates etc. Entities in the DBpedia ontology have also been typed, making it possible to download datasets that are specifically focused on people or geographic locations.

DBpedia has become an extremely popular resource for annotating entities, perhaps in part due to the availability and accessibility of DBpedia Spotlight and an EL service [69, 19]. Even with the general apprehension around Wikipedia as a source of information among the humanities scholarly community it is extremely common to see DBpedia used as a means of annotating a CH collection [35, 116]. DBpedia is also well integrated with other KB resources as it maintains equivalence links with sources such as GeoNames and YAGO.

Two resources that are similar to DBpedia but were not used in the course of this research are YAGO [108, 53] and Freebase [8]. The YAGO ontology incorporates information from Wikipedia, WordNet and GeoNames and is maintained by the Max-Planck-Institut. The information in the ontology is manually checked for accuracy with confidence values being assigned to all relations. Freebase was a community constructed ontology which was intended to be a repository of knowledge akin to Wikipedia. The Freebase API was officially deprecated in 2016 and, while the final state of the ontology is still available for download it is no longer actively maintained.

2.6.2 Wikidata

Wikidata is a community constructed repository of information that is maintained by the Wikimedia Foundation [114]. It is essentially a Wikipedia-like resource that is more amenable to analysis with computer software. Essentially, the problem with Wikipedia is that, in spite of the vast quantity of information available in its articles, these articles are written to be human readable rather than machine readable. DBpedia's extraction framework can extract information from Wikipedia articles, but with Wikidata, the project is consciously organised to be machine readable.

As with Wikipedia, the content in Wikidata is added to, and maintained by a community of contributors who continuously update and refine the information available. One of the advantages of this model is that, as gaps in Wikidata's content are identified, it is possible to update the KB source with missing information. However, from the perspective of historical scholarly research, this open community contribution is not necessarily a good thing. Even though it is a requirement that the Wikidata community

[3] `https://github.com/dbpedia/extraction-framework`

provide citations for any claims that are added to the ontology, this does not lend adequate credibility to the information from the perspective of a historian.

2.6.3 GeoNames

GeoNames[4] is a database of information about geographic locations in all countries across the globe. The dataset is created by a community of contributors who can add or update locations within the dataset, offering information about points of interest or noting various alternative names by which a location may be known.

GeoNames datasets can be downloaded on a per-country basis, making it extremely easy to download a subset of the repository which pertains specifically to Ireland. The Irish dataset contains information about 26,457 unique locations in the Republic of Ireland.

Ireland is a somewhat problematic case for geographic data sources such as GeoNames due to the separation of the island into Northern Ireland and the Republic of Ireland. The Irish border was not instantiated until 1922 after the establishment of the Irish Free State. Therefore, primary source historical records pertaining to Ireland before 1922 do not distinguish between locations in the North and locations in the South. Consequently, cultural heritage collections which pertain to the island as a whole can be packaged as "Irish" collections when, given modern geographic bounds, they are arguably the cultures of two separate countries.

Further problems arise due to the fact that modern sources of geographic information group the six Northern Irish counties into UK datasets while the remaining 26 counties exist in their own Irish dataset. Such is the case with GeoNames. Although awkward, this is not necessarily a serious obstacle. Both datasets may be downloaded and used to annotate a target collection. If there is a need to remove locations that exist in England, Scotland and Wales then the UK dataset can be filtered to only include locations within a bounding polygon around the island of Ireland.

A KB built on GeoNames can be integrated with DBpedia through DBpedia's GeoNames Links dataset[5].

2.6.4 Geohive

While GeoNames is a well established source of geographical data, Geohive [21] is arguably a more appropriate ontology for the annotation of Irish collections. This is because the data in Geohive is gathered by Ordnance Survey Ireland (OSI), which is the Republic of Ireland's national mapping agency.

Geohive is considerably more detailed than GeoNames with respect to locations around Ireland, documenting 50,607 townlands, parishes, baronies, and counties in the Republic of Ireland alone. However, Geohive has some limitations when compared with GeoNames.

[4] https://www.geonames.org/
[5] https://wiki.dbpedia.org/downloads-2016-10

The most serious limitation of Geohive is that it omits the 6 Northern Ireland counties. The equivalent for OSI in Northern Ireland is the Land and Property Service (LPS) which encapsulates Ordnance Survey Northern Ireland (OSNI). It is the duty of this mapping agency to document the northern counties in detail. While LPS does provide publicly available information about locations in Northern Ireland, to date this data does not have any form of semantic representation. This does not mean that it cannot be used to construct a KB, although as previously mentioned there is a general assumption in EL that a KB source is semantic.

Consequently, by itself, Geohive does not have sufficient coverage for the types of Irish cultural heritage collections studied during this PhD.

A second limitation is that Geohive does not contain references to any external ontologies such as DBpedia. Conversely, resources such as DBpedia do not link to Geohive. This makes it difficult to create links between collections annotated with Geohive and those that use other vocabularies. Furthermore it creates problems when attempting to perform EL using the methods documented in Chapter 5.

2.6.5 The Irish Genealogies Ontology

An interesting ontology which is under active development and has not yet been peer reviewed is the Irish Genealogies Ontology (Irish-Gen)[6] developed by Christopher Yocum. Although this resource has not yet been formally published, it is an exciting resource which may be very valuable for future EL projects on Irish collections.

Irish-Gen is an ontology of familial connections between Irish historical figures derived from primary source records currently curated on the Corpus of Electronic Texts (CELT)[7]. The ontology was automatically generated using Perl scripts which were executed over digitised and TEI annotated manuscripts available on CELT. The resulting RDF files are currently undergoing a manual correction process to remove parsing errors.

The genealogies documented in Irish-Gen describe familial relationships in medieval Ireland (typically circa the 12^{th} century). Resources used in the construction of the ontology include the Book of Leinster, The Laud Genealogies and The Book of Glendalough. As with any historical resource, there is a distinction to be made between what is factual and what is recorded in these documents. For example, if an individual was sufficiently wealthy to commission an genealogy in the 12^{th} century, then it is generally accepted that the scribes tasked with researching the family tree had an unwritten responsibility to ensure that the lineage of their patron eventually linked to a great historical king, hero or other figure of note. Primary source historical documents are rife with issues such as this.

At present, the ontology does not contain information such as birth or death dates which make it difficult to filter the KB to a specific time period. The relationships between individuals are purely familial and do

[6] https://github.com/cyocum/irish-gen
[7] https://celt.ucc.ie/index.html

not capture interactions such as financial transactions, altercations, etc. which may be helpful to the EL process. The project is essentially a semantic family tree. Additionally the ontology in its present form is a stand-alone project that does not make reference to any other formal knowledge bases. With specific regard to the research in this thesis, it is unfortunate that the focus of Irish-Gen is several centuries too early to be of use.

Even so, Irish-Gen is an extremely exciting body of work and should be considered for future work on Irish collections. As it stands, Irish-Gen is the only resource of its kind that has been identified in the course of this PhD.

2.7 Candidate Selection

Given an entity mention as input, the candidate selection process identifies a set of candidate referents in the KB to which the mention may be referring. In many EL systems, this is achieved using simple IR approaches with Lucene commonly being employed to index the KB and mentions being executed as queries against said index [79, 9]. Alternatively, a simple dictionary lookup may be used [116].

There is a balance to be struck between precision and recall when retrieving candidates from the KB (See Section 2.9.3 for a discussion of these metrics in the context of EL). Although it might naïvely be thought that the candidate selection process should retrieve all candidates that have an association with a given surface form, and that a good referent selection process should cut through ambiguity and accurately identify an appropriate referent for a given mention from a large pool of candidates, over-emphasis on recall should actually be avoided. For example, if an EL system uses the relationships between entities to choose correct referents, then it can be difficult to identify meaningful relationships if the candidate referent sets are extremely large.

More practically, larger pools of candidates require greater computational time on the part of the referent selection process in order to arrive at a final result. Irrespective of how good the referent selection process is, working through large sets of candidate referents can be the difference between an EL task that takes seconds versus one that takes minutes, sometimes with little to no difference in the quality of the output.

The candidate selection process may screen the pool of candidates in order to remove those that are obviously unsuitable based on some suitability measure. Trivially, filtering may be done using string similarity between labels associated with the candidate in the KB and the spelling of the mention. A threshold similarity is applied and candidates which fall below this threshold are removed from the pool.

Alternatively, the types of entities may also be considered as criteria for inclusion in the candidate pool. KEA, for example, performs NERC before beginning the process of EL [96]. When candidates are retrieved from the KB they are filtered so that the type of all candidates must match the entity type as determined during NERC. Essentially, there is no point in considering Paris Hilton as a referent for the mention “Paris” if a NERC tool has already determined that the mention must be referring to a place

based on the surrounding context.

Depending on how the surface forms in the KB are sourced, some approaches to candidate selection may consider the probability that a surface form is a link to a given entity. Medelyan et al. [68] for example computes the probability that a given surface form, when used as anchor text, maps to a specific instance of an entity (this measure is known as *commonness* in the original publication).

Although this approach was tested in the context of Wikification, information about link probabilities can still be derived from KB sources such as DBpedia which maintains an anchor text dataset. Using this probability measure, a cut-off may be established for a candidate whereby the probability that the surface form is a reference to the candidate is too low relative to all other candidates being considered.

Of course, the most effective method of ensuring that an inappropriate candidate is not considered as a referent for a mention is to simply not index the candidate in the KB. If an EL system is deployed for use on a collection of geographic data, then there is no need for the KB to contain URIs for people. Approaches which attempt to apply automatic enrichment processes to CH data will often filter the content of the KB to remove any entities that are guaranteed to be irrelevant [102]

2.8 Referent Selection

Given a pool of candidates from which to choose a referent for each mention the remaining challenge is to select a single referent for each mention, including the special NIL annotation, should no suitable referent exist. In order to describe this process, this thesis will adopt the formalisation of EL presented by Ratinov [93] who expressed the challenge as an optimisation problem of two parts – global and local similarity.

Let M be defined as the set of all entity mentions in document d (where d could be a book, a chapter, a paragraph, a sentence, etc). Let m_i be an individual entity mention such that $m_i \in M$. Let C be the set of all candidate entities that are present in the KB. For each mention m_i, a set $C_i \subseteq C$ is generated where each entry $c_{i,j} \in C_i$ is a candidate referent entity for m_i. Let Γ be an output set of size $|M|$ where $\Gamma \subseteq C \cup \{\text{NIL}\}$ and $\gamma_i \in \Gamma$ is the chosen referent for m_i. It should also be noted that $\gamma_i \in C_i \cup \{\text{NIL}\}$.

Given these definitions, Ratinov defines EL to be a maximisation problem of the following form:

$$\Gamma^* = \underset{\Gamma}{\operatorname{argmax}} \sum_{i=1}^{|M|} \big[\phi(m_i, \gamma_i) + \psi(\Gamma)\big] \tag{2.1}$$

In this definition ϕ is a local similarity function which establishes the strength of the relationship between m_i and $c_{i,j}$ in a manner that is not dependent on the similarity between any other mention-entity pair. In other words $\phi(m_i, c_{i,j})$ is independent of $\phi(m_x, c_{x,y})$. Implementing ϕ, may be as trivial as computing

Symbol	Description
d	The document whose ambiguous entity mentions are to be disambiguated
M	Set of ambiguous entity mentions found in d
m_i	An individual ambiguous mention $m_i \in M$
C	Set of all candidate entities in Knowledge Base
C_i	Set of candidate entities to which m_i might be referring $C_i \subseteq C$
$c_{i,j}$	An individual candidate referent entity $c_{i,j} \in C_i$
NIL	A special value indicating that no referent entity exists for a mention NIL $\notin C$
Γ	The set of referent entities identified for each mention in M . $\Gamma \subseteq C \cup \{\text{NIL}\}$
Γ^*	The best permutation of Γ as determined by some scoring function
γ_i	The chosen referent entity for m_i. $\gamma_i \in \Gamma$ and $\gamma_i \in C_i \cup \{\text{NIL}\}$
ϕ	Local Similarity Function
ψ	Global Coherence Function

Table 2.1: A summary of the symbols introduced so far

the textual similarity between the surface form of m_i and surface forms which are known to refer to $c_{i,j}$, in which Jaro-Winkler similarity [118] or Levenshtein distance [65] between surface forms may be used.

ψ is a global coherence function which measures the homogeneity of Γ with respect to some overarching subject matter. The inclusion of ψ is based on the idea that multiple entities extracted from the same document are likely to be related to the same (or at least similar) topics. If some relationship can be established between the disambiguation candidates, then the obvious candidates can help to discern the referent entity for more ambiguous candidates. In the crudest sense, this could be though of as permuting all possible values of Γ and calculating some agreement between referents, with the objective of finding the permutation with the highest agreement. Precisely how "agreement" varies from implementation to implementation.

It is generally accepted by the community that finding a solution to ψ in this model is an NP-hard problem [93, 78]. Hence EL systems constructed based on Ratinov's model often propose approximations or substitutes for ψ.

The task is to identify Γ^*, the configuration of Γ which maximises both $\phi(m_i, \gamma_i)$, and $\psi(\Gamma)$. Ostensibly Γ^* will be the optimal configuration of referent entities for M.

Truthfully, this formalisation of the problem is not universal. Some researchers actually view the problem as a minimisation problem [124] while others compute the final score for a set of assignments as a product of the function outputs, rather than the sum [44]. Some approaches even switch between formalisations

depending on the nature of the source text [3].

There are a multitude of approaches which attempt to approximate a solution to this formulation of the EL problem. An excellent overview of these solutions has been provided by Shen et al. [101]. Broadly these methods of approximation can be categorised into three classes:

1. Probability priors
2. Coherence measures
3. Contextual similarity

2.8.1 Probability Priors

Probability priors are statistical measures obtained from analysing the KB. These priors help to weight certain decisions in the choice of a referent. Many simple probability priors are seen in Wikification, but these can translate to more KB agnostic approaches. For example, the previously mentioned commonness measure by Medelyan et al. [68].

Commonness is a probability prior which is simply the likelihood that a particular surface form refers to an entity. It is computed based on how often the surface form points to a certain Wikipedia article whenever the surface form is used as anchor text. For example, the anchor text "Dr. Henry Jones" will usually link to an article about a character from the 1989 film "Indiana Jones and the Last Crusade"[8], but occasionally it will instead link to the article for the 17th century Anglican bishop[9]. In the absence of all other evidence, selecting the most common referent for a surface form is a reasonably sensible decision.

Popularity of an entity is another extremely cheap and simple probability prior that may be computed. Historically this was computed based on user interactions with Wikipedia. The popularity of an entity in the KB was determined by how often its Wikipedia article was viewed.

A more modern method of computing probability priors is used by the EL system Probabilistic Bag of Hyperlinks (PBOH). PBOH learns a probability distribution which determines the likelihood of a candidate being the correct referent 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 [40]. This problem is NP-hard, hence the resulting probabilities are approximated in practice using loopy belief propagation [80].

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

[8] https://en.wikipedia.org/wiki/Henry_Jones,_Sr.
[9] https://en.wikipedia.org/wiki/Henry_Jones_(bishop)

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.

2.8.2 Coherence Measures

Coherence measures measure the level of agreement between a given subset of entities. The theory is based on the assumption that entities in a document are likely to be related by some common topic. Hence the correct referents for the mentions will be identified based on some mutual relationship which they all share.

A popular approach to measuring coherence is to use graphs. Theoretically, entities in the KB which share some underlying relationship are likely to reference each other in some way. By using candidate referents as seeds and growing a graph, it is possible that candidates which are related by a common topic will quickly become linked. These links will form dense clusters in the graph which could be used to assign a rank to the candidates of more ambiguous surface forms.

Hachey [44] tested several graph weighting measures including *PageRank* [85] and *Degree Centrality* [37] to assign a weight to each candidate in the subgraph. Usbeck [112] used Hyperlink-Induced Topic Search (HITS) and averaged the hub and authority scores of vertices in order to identify referents. Numerous approaches have employed this approach of generating a graph and identifying authorities therein [3, 9, 20, 124, 43].

Precisely how the graph is constructed and weighted is entirely application dependent. For example, Babelfy [78] computes a set of semantic signatures for all concepts present in the knowledge base during a pre-processing stage. 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. A completely different approach that is still based on graphs is used by KEA which has been documented in Section 2.8.4.

Alternatively, a coherence measure may examine the semantic similarity of entities and assume that the correct referents are those that are most semantically similar. Milne and Witten used Wikipedia Link Based Measure (WLM) [119] to compute semantic similarity. The measure is defined by the formula:

$$WLM(a,b) = \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))} \tag{2.2}$$

where A and B are the sets of incoming links for the Wikipedia articles which correspond to the candidate referents a and b. W is the set of all links in Wikipedia. The intuition behind this measure is that entities which are semantically similar will share many of the same incoming (or outgoing) links. By measuring

the overlap between the two sets of links relative to the scale of Wikipedia it should be possible to discern a rough estimate of the semantic similarity between a and b.

AGDISTIS is a graph-based EL system which uses the well known HITS algorithm to select referent entities [112]. A set of candidates is retrieved from the knowledge base and a Breadth First Search (BFS) 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.

A related EL system is the Federated knOwledge eXtraction Framework (FOX) which can be used to perform both NER and EL on an input corpus. FOX uses AGDISTIS to perform EL [105].

2.8.3 Contextual Measures

Contextual measures try to use the context from which an entity mention was extracted to identify a referent in the KB. These approaches assume that it is possible to model the expected context for all entities in the KB. For EL systems that use some derivative of Wikipedia e.g. DBpedia as the KB, this is a safe assumption as the textual content of Wikipedia articles provides long form descriptions of all entities. By extracting key terms from articles or building some contextual model based on the language of articles, it becomes possible to compare the context of the mention with the expected context of all its candidate referents.

DBpedia Spotlight [69] is an example of an EL system which uses context to identify a referent for a mention. This EL service uses a Vector Space Model (VSM) to choose an appropriate referent for each mention. Every entity in the KB is assigned a contextual description comprised of the concatenation of all paragraphs that reference the entity in Wikipedia. Spotlight 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 IR problem. The similarity between an input mention (and the contextual text surrounding the mention) is compared with the context all of its candidate referents using cosine similarity. The candidate with the highest similarity score is chosen as the referent.

There are, in fact, two different versions of DBpedia Spotlight [19], the second of which uses a generative model developed by Han et al. [46] to compute the probability that a candidate referent is the correct referent for a given mention. This model arrives at a score for a referent based on a combination of three probabilities:

1. The probability that a link in Wikipedia links to the Wikipedia article for a candidate
2. The probability that the text of the entity mention is used as anchor text in a link to the candidate referent
3. The probability of arriving at the context of the mention given a language model for the candidate referent

The language model is a smoothed unigram model derived from Han's paper [46].

Interestingly, both of the Spotlight approaches only ever consider the local similarity function from Ratinov's formalisation. The relationships between candidate referents are never considered, making DBpedia spotlight a local disambiguation system.

Approaches such as that used by Zwickelbauer et al. [124] also use contextual features to select appropriate referents for mentions. However, their approach also considers the relationship between entities. For Zwickelbauer, word embeddings are used to determine the context of a mention and compare that with the context of referents in the KB. Where a candidate's expected context is too dissimilar to that of the mention, the candidate is removed from consideration. The remaining candidates are used to seed a graph to which PageRank is applied in order to identify referents.

2.8.4 Further Entity Linking Systems

In addition to the EL systems already described above, this section presents a selection of systems which illustrate how the different referent selection methods that have been outlined may be combined. The referent selection process for a number of EL systems have been described below. Some of these are, in fact Wikifiers, but the methods they use can translate to more general KBs.

KEA

It was previously mentioned in Section 2.7 that KEA uses NERC to identify and classify mentions of entities in free text. When choosing candidates from the KB, the types of the recognised entities are used to filter the set of candidates and remove those that do not match the type assigned by the entity classifier.

Given a set of candidate referents retrieved from the KB, a graph is created based on links between entity articles in DBpedia. Links are only created between candidates which are not competing directly with each other. For example, the entities "Steve Jobs" and "Laurene Powell Jobs" might both be considered as referents for the surface form "Jobs". Even though a relationship between these entities exists, it is not considered in the graph because the two are in direct competition. However, if there is also an entity "Apple Inc." which may be the correct referent for the surface form "Apple", then the link between the "Steve Jobs" candidate and the "Apple Inc." candidate will be considered. A sequence of analysis are then performed on this graph in decreasing order of reliability.

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 [107].

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. An entity is unambiguous when all other competitors for a mention have been removed from consideration. Hence it is possible for the disambiguation process to halt before getting through the full four steps. If no candidate has been found for a mention after executing each of the four steps, KEA applies a NIL label.

2.8.5 WikiMiner

WikiMiner [74, 73] uses features derived from the link structure of Wikipedia to train a simple classifier which can weight candidate referents for entity mentions. The training data is constructed from internal links between articles in Wikipedia. Based on this data, Milne and Witten defined three features by which the quality of a candidate as a referent for a surface form might be ranked: *commonness*, *relatedness* and *goodness*.

Commonness is a probability prior as described in Section 2.8.1.

Relatedness is a global measure which determines how coherent a chosen referent is with respect to the context established by other candidate referents being considered. This is computed as a weighted average of the semantic similarity between a candidate referent and all other candidate referents in the document. Milne and Witten used Wikipedia Link Based Measure (WLM) [119] to compute semantic similarity. Relatedness is thus a weighted average of the scores produced by WLM between a candidate referent and all other candidate referents being considered.

The weights in the averaging process are determined by how often a surface form is seen to refer to an entity in the training data. For example, the word “here” is commonly used as anchor text for a link, which means that WikiMiner will consider it a valid surface form for entities. Yet “here” is also frequently occurs outside anchor tags, meaning that it is not always a reference to an entity. Hence the weight of a candidate referent in the averaging process is based on how often it is used as anchor text with respect to how often it occurs in the training corpus.

Goodness refers to the homogeneity of the context of the document. If the document is reasonably consistent, then relatedness can be a reliable measure of quality for a candidate referent. If the subject matter of the document is more erratic, then it can be useful to weight in favour of commonness. Goodness is computed as the sum of the weights described in the previous paragraph.

Given these three measures, a supervised classifier is trained to score the quality of candidates as referents for a surface form. In its default configuration WikiMiner uses Weka’s J48 Decision Tree [89] for classification.

2.8.6 TagME

TagME [32] was developed as a method of annotating extremely short texts. It uses a vote based scheme whereby, after candidates were identified for each mention, the "goodness" of a candidate for a given are was voted upon by the candidates for all other mentions being considered. This goodness score was computed using WLM [119]. However, the amount of influence that a candidate exerts is determined by how ambiguous the candidate itself is. This is found by computing how often the surface form is associated with the given candidate (essentially, how often the surface form was used as anchor text that linked to the candidate), and how many different candidates the surface form had.

The quality of a candidate was thus determined by the number of votes it received from all other candidates being considered by other mentions. TagME then uses one of two classifiers to arrive at a referent. The first classifier classifier computes a score based on joint probability derived from the ambiguity of the mention and the number of votes that a candidate received. The highest scored candidate according to this classifier is chosen as the referent. Alternatively a threshold based classifier chooses the sense that received the highest number of votes.

After referents have been assigned to all mentions, the set of referents is pruned to remove all those that are not coherent with respect to the rest of the chosen referents.

TagMe has since evolved into TagME 2 [31] and later into WAT [87].

WAT provided several extensions on top of the original TagME. Indeed, WAT is more of an ensemble of approaches to EL than it is a single approach. In its simplest form, WAT is simply TagME with the possibility of adding a string distance measure and/or a weight based on the results of the original TagME scoring function. Context of a mention is now also considered.

WAT also incorporates entity mentions graphs akin to those used in AIDA (described below). However, instead of computing a dense subgraph, WAT applies a graph ranking algorithm such as PageRank or HITS to score nodes in the graph. The initial weights of the graph are computed according to one of three measures: commonness (the number of times the text of the mention was used as anchor text to link to the candidate), context (computed using BM25) or an identity score that was simply 1.

As before, WAT will abstain from supplying a referent for mentions whose highest ranked candidate falls below a given threshold.

2.8.7 Dexter

Dexter [13] is an NER and EL framework which implements three different EL methods from the literature. Two of these have already been discussed, namely TagME and WikiMiner.

The third method is a method known as collective linking [47]. This is a graph-based method which assigns weights to candidate entities based on a combination of importance of the mention to the sur-

rounding context, compatibility of the candidate with the mention and coherence of the candidate with respect to other candidates.

AIDA

AIDA's disambiguation algorithm [120, 52] uses three features in the selection of a referent for each input surface form:

- An entity prior derived from the popularity of the entity in the KB.
- The similarity between the document from which the surface form is extracted and a set of keywords which describe each entity in the KB. In other words, does the context of the surface form match common contexts for any given candidate referent.
- The coherence of any combination of candidate referents. In the paper, WLM is used to determine this score.

AIDA also includes tuning parameters for each of these features which may be scaled up or down in order to lend greater or lesser weight to any individual feature as the problem dictates. Automated robustness tests in AIDA can switch on or off the entity prior or the coherence test if the nature of the surface forms suggests that their inclusion would be detrimental to the overall disambiguation process.

Given these three measures, AIDA generates a mention-entity graph where nodes in the graph are comprised of the surface forms (mentions) and the candidate referents (entities). Weighted edges link mentions to entities, and entities to other entities. The weights on a mention-entity edge are based on the entity prior and contextual similarity. Entity-entity edges are weighted based on the coherence of the entities. AIDA then proceeds to prune this graph in order to determine a dense sub-graph which should indicate the correct referents for each mention.

2.8.8 Babelfy

Babelfy combines the tasks of WSD and EL in order to present a unified method of semantically annotating text. During a the construction of the KB, a set of semantic signatures is generated for all concepts present in the KB source. After these signatures have been generated, an arbitrary input text may be processed for linking.

The signatures were generated by analysing relationships between entities in the KB source. For every cycle of three vertices v_1, v_2 and v_3 such that there existed an edge sequence $v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_1$, the weight of the edge between v_1 and v_2 was incremented.

A random walk with restart [109] is run over the resulting graph with initial edge weights being determined by the previously described directed cycles. The strength of the relationship between a vertex v_1

and a second vertex v_2 is determined by how many times v_2 is encountered during a walk which began at v_1. The semantic signature of the vertex v_1 thus becomes the set of vertices that are most frequently encountered during its walk. There is a cut-off N which is the maximum number of vertices that can be added to a semantic signature.

When linking entities in an input document, a set of candidates is generated for each mention in the text. These candidates are used to seed a graph where each vertex is a tuple of a mention and a candidate. Edges are then created between vertices provided

1. The mention of two vertices is not the same
2. The candidate of two vertices are contained in each others semantic signature

A densest subgraph of this new graph is computed and the chosen referent for a mention is selected as the vertex candidate with the highest score, where a score is computed as the normalised degree weight. A hard threshold is applied which can assign a NIL to a mention if the score of the candidate is too low.

2.8.9 EARL

Entity and Relation Linker (EARL) [24] combines entity linking with relation linking specifically for the purposes of responding to natural language queries expressed by users over a knowledge graph. Whereas EL is the task of identifying a referent for a surface form in a query, Relation Linking involves disambiguating the nature of a relationship expressed between two entities e.g. “Tom Waits *wrote* Rain Dogs”.

EARL’s EL process begins by extracting keywords from a query using SENNA [15], followed by the classification of keywords as denoting either entities or a relationship. This is achieved using an long-short term memory network (LSTM) as a classifier.

As is typical, given each keyword’s clasification as either an entity or a relationship, a search index (Elasticsearch[10]) is used to retrieve candidate URIs for each mention.

At this point, EARL provides two solutions to the linking problem (both Relation and Entity Linking).

The first solution attempts to find the correct referents by modelling disambiguation as an instance of the General Travelling Salesman Problem (GTSP). EARL generates a graph comprised of vertices, edges and edge labels $G = (V, E, L)$. This graph can be partitioned into subgraphs where a subgraph is a group of vertices V' which are all candidate referents for the same mention. A cost function is defined based on the distance between vertices between sub-graphs. The objective is to find a choice of referents which minimise this cost. EARL approximates a solution to this problem using the Lin-Kernighan-Helsgaun algorithm [50].

[10] `https://www.elastic.co/products/elasticsearch`

The second solution treats the linking problem as a ranking problem. The objective is to rank the set of candidate referents in order of suitability as referents for a given mention. This is achieved using Connection Density. Here, the candidate referents are retrieved as before from the search engine, with their initial rank being the rank provided by the search engine i.e. the rank of search results. Again, the candidate referents are grouped so that the set of candidates for a mention belong to the same set.

The number of "hops" is computed between candidates between these sets, where a hop is defined as the shortest path between two candidates based on the number of edges that must be traversed. This yields Hop-Count, which is the sum of the distances from a candidate to all other candidates divided by the number of keywords in the query.

EARL also defined a Connection-Count feature which is the number of connections from a candidate to all other candidates in other sets, divided by the number of keywords in the query.

A classifier is trained on Connection-Count, Hop-Count and the initial rank to determine the probability that a given referent is the correct referent. The original paper found that extreme gradient boosting (xgboost) yielded the best results on their evaluation dataset.

2.8.10 DeepType

DeepType [90] tackles the EL problem by focusing on correctly identifying type labels for entities, rather than developing a sophisticated method of identifying a referent. Indeed, in the original paper simple link probability is used to determine the quality of a candidate as a referent i.e. how often the surface form was used as link text with the given candidate as a target. This link probability is combined with a sophisticated entity classification system (or typing system) to identifying the correct referent. Interestingly this work shows that accurate answers to the entity classification problem can yield high accuracy in the EL problem.

2.9 Evaluating Entity Linking Systems: Challenges and Approaches

Performing a reliable, reproducible evaluation of EL algorithms can be a surprisingly difficult task. This is partially due to the fact that the definition of a "correct" annotation can be somewhat subjective.

For example, assume the existence of an EL service which performs NER and EL in sequence. It is capable of taking raw, unannotated text as input and returning a semantically annotated result as output. The KB is DBpedia.

The service is given the sentence "Prince Rogers Nelson was a musician" to annotate. It annotates the word "Prince" with the URI `dbp:Prince_(musician)`. The system has correctly identified an entity and the corresponding referent, but it has failed to annotate the full span of the entity mention. Should the system be penalised for this? If so, what is an appropriate penalty?

In a follow-up task, the service is tasked with annotating "The Tallest Man on Earth is a musician". The service annotates the surface form "Earth" with the URI `dbp:Earth`. In this instance, the problem with identifying the word boundary results in an incorrect referent being identified. Once again, should the system be penalised for this and to what extent? The mistakes of the EL algorithm are predicated on the mistakes of the entity recognition algorithm. Thus it is perhaps overzealous to condemn the performance of the EL algorithm given that it performed quite sensibly for the mention that was identified.

Alternatively to the word boundary problem, consider equivalence between URIs in semantic resources. When using DBpedia as a KB, redirects may be considered as an expression of equivalence between entities. For example the entities:

`dbp:Prince_Rogers_Nelson`

`dbp:Rogers_Nelson`

`dbp:The_Artist_Formerly_Known_As_Prince`

are all redirects to the entity `dbp:Prince_(musician)`. One might thus assume that an EL system which assigns any of these four URIs to the surface form "Prince Rogers Nelson" has accurately identified the correct referent.

Consider now the DBpedia entity `dbp:PEHDTSCKJBMA`[11], an acronym comprised of the first letters of cities visited by the musician Tom Waits during his 2008 tour: Phoenix, El Paso, Houston, Dallas, Tulsa, St Louis, Columbus, Knoxville, Jacksonville, Mobile, Birmingham and Atlanta[12]. In the DBpedia ontology, this entity is a redirect to Waits himself `dbp:Tom_Waits` despite the fact that, while the acronym has strong associations with Waits, it is not equivalent to him semantically in any real-world sense.

How should an annotation system be scored given semantic problems in the KB? Is an annotator which assigns the URI `dbp:Prince_(musician)` to the surface form "The Artist Formerly Known As Prince" correct? If so, then what of an annotator which assigns the URI `dbp:PEHDTSCKJBMA` to the surface form "Tom Waits"?

Exacerbating the situation is the fact that the performance of a system can vary wildly depending on the nature of the content type that comprises the evaluation corpus. EL on a Twitter corpus is significantly different to EL on a news corpus due to variation in context size, linguistic features and prevalence of entities encountered [95]. Hence if the results of an evaluation of two different EL systems are published based on two different test corpora, it is difficult to say much about their relative performance.

[11] The spelling of the acronym in the DBpedia URI is erroneous. It should be PEHDTSCKJMBA.

[12] Waits has claimed that PEHDTSCKJMBA also holds a deeper meaning, the interpretation of which is left open to discussion: People envy happiness. Dogs though, sense courage knowing jubilation means better assets.

In CH yet another problem faced is the fact that collections may need to be annotated using multiple different vocabularies [67]. Where an entity exists in two different vocabularies e.g. a location that is present in both GeoNames and DBpedia, an evaluation system must allow for this.

Efforts have been made to standardise how EL systems are tested and compared. Precision, Recall and F1 measure from the field of IR are commonly employed, although there can be some variation in the precise definition of Precision and Recall e.g. "pooled recall" and "pooled precision" used by Manguinhas [67] as compared with "Macro" and "Micro" P and R used by the GERBIL platform [111]. Even so, there have been attempts to standardise these evaluation measures for EL in order to introduce better consistency into how systems are tested and compared.

2.9.1 Common Evaluation Metrics

Entity Linking can conceptually be thought of as an IR task. Surface forms are queries and the objective is to always return the correct referent as the top ranked result, or indeed to return no result where the referent is not found in the KB. It is therefore unsurprising that the well established IR metrics, P, R, and $F1$ have become the most common means of assessing and reporting the performance of EL services.

The formulation of these three statistics are well known:

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fn}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

Where tp, fp, tn and fn are counts, the definitions of which are as follows in the context of EL:

- **True Positive:** annotator assigned a correct referent.
- **False Positive:** annotator assigned an incorrect referent.
- **True Negative:** annotator correctly did not assign a referent.
- **False Negative:** annotator incorrectly did not assign a referent.

It can be seen from the definitions of P and R that the value of tn does not impact either of the measures. This count can thus be ignored.

This section has already provided some examples of how subjectivity can creep into EL evaluation. Even with the seemingly reliable P, R and $F1$, there is room for interpretation.

A test corpus is comprised of multiple documents. Each of these documents may contain multiple instances of mentions. Should an EL system be evaluated based on how well it disambiguates all of the mentions across all of the documents, or should its performance be assessed individually for each document and the results averaged?

One approach is to evaluate an EL system based on both of these perspectives and report the results for each. This has led to the definition of two sub-categories of P, R and $F1$, namely micro and macro P, R and $F1$.

Micro considers the entire collection as a single EL problem. The total scores for tp, fp and fn are calculated across the entire collection and used to compute P_{micro} and 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:

$$
\begin{aligned}
P_{micro} &= \frac{\sum_{d \in D} |tp_d|}{\sum_{d \in D} |tp_d| + |fp_d|} \\
R_{micro} &= \frac{\sum_{d \in D} |tp_d|}{\sum_{d \in D} |tp_d| + |fn_d|} \\
F1_{micro} &= \frac{2 \times P_{micro} \times R_{micro}}{P_{micro} + R_{micro}}
\end{aligned}
\tag{2.3}
$$

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, R and $F1$ are calculated for each document using the formulae given in Equation 2.3. The values of P_{macro} and R_{macro} and $F1_{macro}$ are the average of P, R and $F1$ 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:

$$P_{macro} = \frac{\sum_{d \in D} P_d}{|D|}$$

$$R_{macro} = \frac{\sum_{d \in D} R_d}{|D|} \tag{2.4}$$

$$F1_{macro} = \frac{\sum_{d \in D} F1_d}{|D|}$$

What has been outlined so far in this section is essentially the evaluation metrics proposed by the Benchmarking Entity-Annotation Systems (BAT) framework [16] which defines a number of experiment types and performance measures aimed at standardising the evaluation of EL tools. The implementation of these experiment types and metrics have been crystallised in the GERBIL platform [111] which will be discussed in Chapter 3.

2.9.2 Limitations of Evaluation Metrics

A limitation of the evaluation metrics presented in Section 2.9.1 is that they do not provide information on the source of error should an EL service produce incorrect results. This is understandable considering that disambiguation is usually treated as a black-box process. An evaluation corpus is submitted, a set of annotations are retrieved and a direct comparison between actual and expected annotations is performed. However, for such cases where access to the internal machinations of an EL service is possible, it may be insightful to perform a more thorough investigation [83].

For the purposes of this discussion, consider an input document d which contains a mention m_i. The correct referent $\mathcal{G}_i$ is known in advance. An EL service is tasked with finding a referent γ_i which identifies the subject of m_i. The service returns an incorrect annotation $\gamma_i \neq \mathcal{G}_i$, suggesting a problem with the disambiguation algorithm. The problem could stem from one of three scenarios:

1. $\mathcal{G}_i$ is not in the KB, so it is impossible for the service to return the correct disambiguation.
2. $\mathcal{G}_i$ is in the KB, but was not returned by the candidate selection process. The issue lies in how the KB is indexed and searched, or with the inclusion criteria for the candidate set C_i.
3. $\mathcal{G}_i$ was correctly retrieved from the KB and is in C_i. The disambiguation algorithm failed to correctly identify $\mathcal{G}_i$.

The evaluation approach described in Section 2.9.1 conflates these three sources of error which obfuscates the true source of inaccuracy during an experiment. This is a remarkably important consideration which

is easily neglected. In particular, it is easy to ascribe all successes and failures to Scenario 3 when performing a black box evaluation.

A reasonably foolish EL algorithm can yield seemingly accurate results if the candidate selection process is brutal enough in its selection process, or indeed if the scope of the KB is narrow enough. The disambiguation algorithm does not need to be particularly clever if every entity in the KB has a distinct and clearly different name. Keep this in mind when reading Section 2.10, which discusses applications of EL to CH.

The next section discusses some evaluation metrics which can complement those proposed by the BAT Framework.

2.9.3 Additional Evaluation Metrics

Hachey et al. [45] proposed a set of evaluation metrics which in turn are taken from the Text Analysis Conference (TAC) standard evaluation measures. Rather than focusing solely on the accuracy of the annotations, these measures investigate the amount of ambiguity faced by the EL system. By extension they can be said to indicate how well an EL algorithm deals with ambiguity, reflecting an important consideration in EL evaluation.

Candidate Count is simply the average size for each candidate disambiguation set. Intuitively, smaller candidate sets generally mean that there is less ambiguity for the disambiguation algorithm to resolve:

$$\mu_C = \frac{\sum_i |\, C_i \,|}{|\, M \,|} \tag{2.5}$$

Candidate Precision measures the percentage of candidate sets which contain the correct disambiguation for their associated mention under the condition that the candidate set is not empty i.e. only sets where the mention has at least one candidate are considered. The formulation for Candidate Precision may be expressed as:

$$P_C = \frac{|\{C_i \,|\, C_i \neq \varnothing \wedge \mathcal{G}_i \in C_i\}|}{|\{C_i \,|\, C_i \neq \varnothing\}|} \tag{2.6}$$

Candidate Recall is the percentage of candidate disambiguation sets which contain the correct entity disambiguation for their respective mention where the gold standard does not indicate that the system should assign a NIL label:

$$R_C = \frac{|\{C_i \,|\, \mathcal{G}_i \neq \text{NIL} \wedge \mathcal{G}_i \in C_i\}|}{|\{\mathcal{G}_i \,|\, \mathcal{G}_i \neq \text{NIL}\}|} \tag{2.7}$$

As well as determining how well a system performs when identifying candidate entities, it must be remembered that another important behaviour of EL systems is that they should abstain from annotating a mention where no suitable candidate exists. In some cases, this decision may be the result of the candidate selection process not finding any candidates, rather than because the referent selection process rejected all presented candidates. In this case, the candidate entity set will be empty as there are no good candidates to be found in the KB. This results in two precision and recall measures.

Nil Precision is the percentage of candidate sets which correctly contain no candidates:

$$P_{\text{NIL}} = \frac{|\{C_i \,|\, C_i = \varnothing \wedge \mathcal{G} = \text{NIL}\}|}{|\{C_i \,|\, C_i = \varnothing\}|} \tag{2.8}$$

Nil Recall is the percentage of gold standard annotated entities that have been assigned a value of NIL whose corresponding candidate entity set is empty:

$$R_{\text{NIL}} = \frac{|\{C_i \,|\, \mathcal{G}_i = \text{NIL} \wedge C_i = \varnothing\}|}{|\{\mathcal{G}_i \,|\, \mathcal{G}_i = \text{NIL}\}|} \tag{2.9}$$

Hachey also provides a means of evaluating the quality the final chosen referents. This is simply the percentage of entities that were correctly labelled either with an entity label or NIL as appropriate according to the gold standard $\mathcal{G}$. This accuracy score can be computed as follows:

$$A = \frac{|\{\gamma_i \,|\, \gamma_i = \mathcal{G}_i\}|}{|M|} \tag{2.10}$$

A challenge with applying Hachey's measures is that they require that the EL system being investigated exposes the results of its candidate selection process by returning all considered candidates. Failing that, running a local deployment of a given EL system with the source code modified to appropriately log this information is also a possibility.

For the purposes of this thesis, Hachey's measures are used to perform a deeper analysis of an EL system in order to investigate the quality of the candidate selection process. However, due to the difficulty involved in computing these measures where an EL system does not natively support them, this is only done where the results of a GERBIL evaluation would suggest that such an investigation may yield interesting additional insights.

The accuracy score A is substituted with the results output by GERBIL i.e. the evaluation metrics presented in Section 2.9.1 as this breakdown is believed to be more insightful than a single accuracy score.

2.10 Entity Linking in Cultural Heritage

There have been a number of previous efforts to apply EL to existing CH collections. Generally speaking, the challenge presented by a given collection can be measured in terms of the language used (is it a multilingual corpus), the century from which the collection originates (the older a collection is, the more difficult it can be to identify referents), the nature of the content (newspaper articles vs. literary criticisms for example) and the nature of the entities that are the focus of the investigation (often people or places).

In the context of CH the content types encountered can range from noisy primary source content to well-behaved, structured metadata that is manually generated by an annotator.

An example of how trivial some applications might be is the application of EL to a work of popular fiction. Characters in works of fiction tend to have very clear and unique names that makes it easy to identify referents based on surface forms alone (although there may be some challenge in cases of mistaken identity as a literary device). If the work is particularly popular, then it is likely that all of the characters will be documented in some sort of KB source.

Contrast this with a collection of legal documents which describe transgressions between two families. If multiple generations of each family are present, and families followed the tradition of passing names from parent to child, it may become very difficult to establish which member of either family is being discussed at any time. The example of "Charles Coote" is a useful one from the perspective of this research. Charles Coote is the name of at least two generations of Coote, who are mentioned somewhere in legal documentation that was targeted by this work. Establishing which Coote is which and ignoring those later Cootes who were born after the events being studied is part of the challenge that is faced in this thesis.

Some investigations [113, 67] involve the application of off-the-shelf tools such as Open Calais [13], DBpedia Spotlight [69, 19] and, more historically, Alchemy API. Others investigate the construction of new EL systems targeted specifically at collections that are the subject of an investigation [116, 9]. Bespoke systems are often quite simple in their design, usually involving little more than a dictionary lookup which maps surface forms to URIs, followed by some simple metric to eliminate undesirable candidates.

For example, Max De Wilde [116] used a dictionary which mapped string literals to DBpedia URIs to perform EL with respect to places in a corpus derived from the Historische Kranten project. The language of the documents was in French and Dutch and the subject matter was comprised of news text obtained from 19^{th} and 20^{th} century newspapers. Candidate entities were retrieved from the dictionary using a lookup. Each candidate was queried against the DBpedia SPARQL API to determine whether or not it was associated with the type `dbo:Place`. If so then the candidate was assigned to the mention. If there were multiple candidates for a single mention that were marked as places in the DBpedia ontology then the longest match was selected as the referent.

[13] `http://www.opencalais.com/`

De Wilde's approach was incredibly simple, but managed to achieve F1 scores as high as 0.633 when evaluated with the neleval tool[14]. However, this method seems to have favoured French text as the scores for the Dutch locations reached no higher than an F1 score of 0.309. Even so, given the simplicity of De Wilde's approach, the overall performance is quite impressive.

Although it has been stated that Wikification and EL are separate tasks, given the strong relationship between the two it is worth noting that Fernando et al. [29] have previously investigated the application of Wikification to CH material. Using WikiMiner [74] the team investigated the quality of annotations applied to a corpus of 21 items comprised of 381 entities obtained from Europeana (see Section 2.10.2 for an introduction to Europeana). The subject matter of these items appears to have been quite diverse, with topics ranging from science and mathematics to humanities and the arts.

Three different approaches were tested using Wikipedia categories (two different approaches) and the link structure of Wikipedia. Of the methods tested the approach using link structure performed the best with an F1 score of 0.684, suggesting that examining the relationships between entities was a beneficial consideration when linking entities on this content type.

Seth Van Hooland [113] performed an investigation which used 378 records obtained from the Smithsonian Cooper-Hewitt National Design Museum. Van Hooland's investigation examined annotations applied to meta-data fields associated with items in the collection, rather than performing annotations directly on the content of an item. Three third-party EL tools were considered: DBpedia Spotlight, Zemanta and AlchemyAPI. Unfortunately this paper lacks the kind of EL evaluation that would be comparable with the previous studies. While there is a solid breakdown of performance in the NER task, detail on the EL task is lacking. However, as a study, this work helps to provide yet another example of the kind of content that Digital Humanities (DH) researchers may be interested in annotating with EL services.

2.10.1 Semi-Automatic Approaches

One perspective on applying semantic annotations to CH collections that is worth mentioning is the modification of EL services to actively include a human in the linking process [103, 48]. In this scenario, the referent selection process is removed from the EL system and the set of candidate referents is returned to an annotator via some suitable user interface. The scholar chooses what they believe to be the correct referent from a list of options provided by the service.

Clearly this is not true EL, and publications on the topic do not describe their solutions as EL solutions. Even so, it is included here as a worthwhile analogue/parallel to the EL approaches that are the major focus of this research.

Including a human in the linking process makes sense as it eliminates some of the mysticism behind EL

[14] https://github.com/wikilinks/neleval

for researchers who may not have a particularly technical background. Given the trepidation with which some historical researchers approach automation, this may help to encourage greater confidence in the kinds of tools that computer science researchers can provide.

A semi-automatic approach does not always scale to larger collections. Efforts to apply this approach to large scale collections such as the vast array of data hosted by Europeana may involve some crowd sourcing effort [48]. However, it is not always possible to simply wave aside the cost of manual effort as being solvable with the input of enough interested members of the public.

Historical scholars invest vast amounts of time into the analysis of the collections that they research. They painstakingly transcribe, annotate, organise, cross-reference and document the sources that they study. To some degree, this gives them a monopoly over the data that is ultimately produced as a result of their efforts. Crowd-sourcing annotations for these researchers is not often a feasible solution purely due to the amount of personal investment that goes into creating the collection in the first place.

Incidentally, this protectiveness also extends to the semantic source that is used to annotate a collection. Where does the information in the source come from? How accurate is it? Frequent allusion has been made throughout this chapter to the "unsuitability" of certain KB sources in this context. Scholars need to know that they can trust the semantic annotations that have been applied to their collection. By extension, they need to know that they can rely upon the accuracy of the information that subsequently follows those annotations. Chapter 4 talks about this problem at some length, but this is a serious consideration and one which needs to be talked about if computer scientists are to be genuinely helpful to CH scholars.

2.10.2 Automatic Content Enrichment in Europeana

Europeana[15] is a major curator of European CH content. At the time of writing, the platform hosts almost 58,000,000 digital artifacts from approximately 3,500 museums, gallaries, libraries and archives around Europe[16]. These are made available through 27 different langauges.

Europeana have conducted some interesting research into the automatic enrichment of the data that they house. The linking methods pursued by Europeana are broader than EL as they include WSD as well. For this reason, this section will use Europeana's terminology "semantic enrichment" or simply "enrichment" to refer to the linking process. Semantic enrichment in Europeana is performed using Europeana's own semantic enrichment framework[17].

Enrichment in Europeana focuses on four different fields associated with an item: places, concepts, agents and time periods. For each field, a different vocabulary is used to perform the enrichment. Places are annotated using GeoNames [41]. Concepts and Agents use DBpedia [64]. Historically GEMET[18] was

[15] `https://www.europeana.eu/portal/ga`
[16] https://pro.europeana.eu/page/reasons-to-share-your-data-on-europeana-collections
[17] `https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/`
[18] `http://www.eionet.europa.eu/gemet/`

also used for annotating concepts, but Europeana has since moved away from the use of this vocabulary[19]. Time periods are annotated using Semium Time[20]. These vocabularies are filtered down to contain a core subset of entities that are relevant to Europeana's content.

Matching from Europeana content to the appropriate vocabulary is achieved using a series of rules, the nature of which depends on which field type is being annotated. For most field types, the choice of a URI is based on a case insensitive match between surface forms in the KB and mentions in the meta-data field. In the case of the names of people, some pre-processing is performed to remove dates etc. from the name string. However it has been found that the enrichment framework can experience problems when dealing with ambiguity in mentions [56].

Europeana have also performed some interesting investigations into automatic semantic enrichment. In 2014 a EuropeanaTech task force set out to investigate the effectiveness of a variety of different content enrichment tools on a subset of Europeana's data [56]. Six annotators were investigated: Europeana's enrichment framework, the European Library's enrichment framework, the Background Link (BgLink) and VocMatch services from the LoCloud project [30], Pelagios' geographic EL tool Recogito [102], and two different EL setups provided by Ontotext which used GATE[21].

Most of the tested services use extremely simple methods of performing EL. VocMatch, Recogito, the European Library, and Europeana's enrichment framework all essentially use a dictionary lookup which maps surface forms to URIs obtained from a controlled vocabulary. BgLink uses DBpedia Spotlight to perform NER and EL. Ontotext uses rules written in JAPE[22] to select referent entities from a KB. The investigation also involved the creation of a gold standard which was ultimately comprised of 1757 annotated items. The corpus was created by first tasking the six participants in the experiments with annotating a subset of 17,300 items obtained from Europeana. A random subset was taken from these 17,300 items and manually checked for accuracy by a team of annotators. Rather than correcting erroneous annotations, it seems that the team was instructed to mark incorrect annotations and these mistakes were ultimately removed from the corpus.

Tools performed both NER and EL in the course of the evaluation. The final performance evaluation of the tools allowed for both strict and flexible boundary matching on entities in the corpus. For the purposes of discussion the statistics discussed below are from the flexible boundary matching evaluation.

All of the tools in the evaluation achieved extremely high precision scores, with the greatest performance being Europeana's enrichment framework at $P = 0.985$ and the lowest being VocMatch at $P = 0.774$. This suggests that where the services chose to provide a URI for a mention, they usually provided the correct referent, with Europeana's framework being almost perfect at identifying the referent.

[19] https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y/edit

[20] http://semium.org/time.html

[21] https://gate.ac.uk/

[22] https://gate.ac.uk/sale/tao/splitch8.html

The recall of all systems was much lower, showing that the systems tended to miss instances of entities which should have been annotated. This is a problem with NER rather than EL, but it impacts the final F1 scores with the best performing system being OntoText at $F1 = 0.576$ and the worst being VocMatch at $F1 = 0.091$. No individual annotation system managed to recognise more than 39% of the mentions in the corpus.

The investigators themselves acknowledge, however, that it is difficult to perform a fair comparison between the tools involved in this study. Tools such as Recogito are specifically designed to link against a KB of geographic data, yet the experiment involved a range of content types across a spectrum of subjects including mathematics, art, and medicine.

This suggests a noteworthy characteristic of all systems that were tested. These EL systems are highly conservative in their approach to annotating content. They have a restricted, tight vocabulary with which they work and seemingly focus on entities that are relatively unambiguous. At least, the focus on low ambiguity entities is suggested by the simplistic dictionary lookup that many of the tools use for referent selection. This caution is important when dealing with CH content.

2.11 Referencement et Desambiguisation d'Entités Nommées

Of particular note when discussing methods of performing EL in a CH context is Referencement et Desambiguisation d'Entités Nommées (REDEN), an EL tool developed by Carmen Brando and Francesca Frontini [38, 39, 9] which implements an extremely interesting approach to EL on niche CH collections. It was developed to tackle the problem of linking entities in French essays and literary works from the 19^{th} century which were being studied by Labex OBVIL[23].

With respect to entity coverage, the 19^{th} century works at the centre of Brando's research suffered from similar problems to that of the 17^{th} century works that were the initial focus of this thesis. Brando observed that this could possibly be remedied by employing more domain specific KB sources. However, information in these more specific resources was potentially too sparse to adequately inform an EL system. Hence Referencement et Desambiguisation d'Entités Nommées (REDEN) was developed to perform EL across multiple KBs, with the domain specific resources providing information pertinent to the collection while more general KBs could provide additional supporting information that may help with the selection of a referent. This, of course, requires that the different KB resources reference each other in some way.

This section will describe REDEN and its unique approach to EL.

[23] http://obvil.sorbonne-universite.site/

2.11.1 Indexing Process: Building the Knowledge Base

REDEN creates a KB using a dictionary of surface forms which maps strings of surface forms to lists of URIs which may denote referents. Functionality is provided to construct these dictionaries automatically from BnF, getty, BnE, LGD and the French DBpedia dataset. If this particular means of constructing a KB is used, then REDEN will apply various transformations to the names of entities which correspond to people in the KBs. Using simple pattern matching, REDEN will identify honorifics, titles, forenames, and surnames present in an entity's surface form. It will then genereate possible alternative surface forms by permuting combinations of titles, honorifics and name parts as well as initialling forenames.

Due to REDEN's strict approach to candidate selection (discussed in Section 2.11.2), this is intended to increase the range of surface forms associated with an entity, making it easier to identify possible referents in the KB

However, when configuring REDEN to work with custom KBs, the generation of these dictionaries must be performed manually, and any transformations which may be applied to surface forms must be defined by the individual creating the KB.

After the dictionaries have been created, REDEN uses them to create a Lucene search index which increases the speed at which candidate referents can be identified at run-time.

2.11.2 Candidate Retrieval

On initialisation, REDEN is configured to use some subset of available KBs as "reference" KBs. The URIs of candidate referents can only be retrieved from these sources. For example, if REDEN was performing EL on a literary work using DBpedia and BnF as KBs but with BnF chosen as the reference KB, then the candidate selection process would only select candidates from BnF. These URIs are retrieved using the index whose construction was described previously in Section 2.11.1. With candidates selected from the reference KB, REDEN retrieves their corresponding equivalent URIs from all other KBs. In its present form, REDEN can only use one reference KB at a time.

There is no fuzziness to REDEN's retrieval process. Candidate referents in the KB must be associated with a surface form that is an exact (albeit case insensitive) match for the spelling of the mention. This is simply a design decision made when implementing REDEN and there is no reason why this hard match cannot be replaced with something more lenient.

2.11.3 Referent Selection

After the URIs for all candidates have been retrieved, REDEN loads their corresponding RDF graphs from each URI's respective KB. This essentially retrieves all outbound edges from the candidate referent in the disambiguation graph. The result is multiple graphs per candidate for an individual mention. REDEN fuses these graphs by merging candidate vertices which are known to be equivalent.

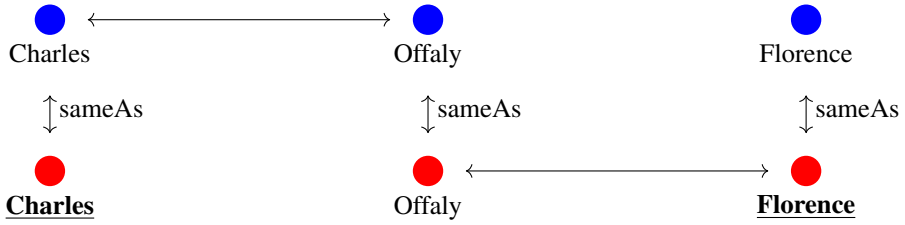


Figure 3: Illustrated example resolving information across multiple KBs

In Figure 3 this merging process is illustrated using two KB sources, denoted by coloured nodes i.e. a red KB source and a blue KB source. The red "Charles" and red "Florence" are being considered as referents for a mention, denoted by the fact that they are underlined. There is a connection between the two entities via the "Offaly" entity. REDEN, during its referent selection phase will download the RDF representation of Charles and Florence from both the blue and red KB sources. It will then merge the red and blue Charles nodes, and the red and blue Florence nodes in an effort to capture information about these entities that is expressed in both KB sources. REDEN will not, however, merge the Offaly nodes, in spite of the fact that an equivalence is indiciated by the KB sources via the sameAs predicate.

The graphs are harvested live at run-time from API endpoints that are known for each available KB. Hence REDEN will only download the RDF representation of entities that are relevant to the problem it is currently solving. However, these KBs can be cached locally, allowing REDEN to run offline and facilitating the annotation of collections with KBs that have not yet been deployed in the Internet.

Once the graphs are fused, REDEN prunes edges and vertices that are deemed unhelpful to the EL process. Any non-candidate vertex which does not have at least two incident candidate vertices connected via predicate edges is deleted from the graph. So REDEN tries to reduce its disambiguation graph to a representation which ensures that any two candidates are, at most, two hops from each other.

REDEN uses a degree centrality measure [36] to weight vertices in the final fused and simplified graph. Multiple different weighting schemes are available, but degree centrality has been shown to be the most effective of those tested in their previous work. For each mention, the candidate with the greatest weight after computing degree centrality is returned as the chosen referent. Mentions with no candidates receive a NIL label.

2.12 Summary

This chapter has introduced the current state of the art in the field of Entity Linking, methods of evaluating EL systems, and the applications of EL to the annotation of CH resources. It is clear that EL is a vast topic, yet the solutions that are typically employed in CH settings tend to be quite simple. The desire has been expressed among the DH community for more rigorous investigations into the applications of EL tools to CH problems, as well as expressing the need for more datasets to support this research.

Of the tools examined in the literature, REDEN presents the most interesting approach withing the context of this thesis' research goals, albeit in a form that is not yet applicable to the problems faced by Irish scholars.

Chapter 3

Evaluating Entity Linking Systems

"Tell me I'm good. Tell me it was good. Tell me I'm good. Tell me that was good. Tell me. I'm good. Tell me I'm good. Tell me I'm good..."
— Deandra Reynolds, *It's Always Sunny in Philadelphia*

3.1 Overview

This chapter presents an evaluation of the performance of 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 3.2, makes them difficult for expert scholars to explore and novice scholars to understand.

The objective is 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 is technically a two part problem as it is important to determine both how well a system can choose an appropriate referent given a mention, and how well the system can abstain from annotating an entity if no suitable referent exists in the KB. In addition to determining how well off-the-shelf EL tools perform at annotating this challenging content, it is also an attempt to identify some desirable implementation details about EL systems.

The investigation is performed by first manually annotating a subset of the depositions with referent URIs taken from DBpedia. It is observed 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 KB provides insufficient coverage for the chosen collection. This is a common problem observed when working with EL systems in CH [113, 1]. The performance of various EL systems is assessed with respect to this ground truth using GERBIL [111], which is a standard benchmarking tool based on the BAT framework [16].

Particular focus is given to two EL systems: Probabilistic Bag of Hyperlinks (PBOH) and AGDISTIS. This was, in part, a practical decision based on observations about nature of sources of data that could be used to inform the EL system (see Chapter 4.1 for detail). One of the assumptions of EL systems which use contextual features to choose a referent is that a long form description of an entity is available which indicates the kind of langauge that can be expected in the vicinity of a mentino. Realistically, when dealing with historical sources, we will have little available information, either in the text being linked or and KB source we may avail of. We can, however, use co-occurrence as a crude indicator of relationships. Hence graph based measures are likely to be more useful for EL problems in CH than those that rely on language models or similar. PBOH and AGDISTIS were two EL systems which relied solely on relationships when choosing a referent.

The results of the evaluation indicate that AGDISTIS implements the most appropriate method for applying EL to the depositions. A deeper analysis of AGDISTIS is performed including a further evaluation using Hachey's proposed EL evaluation measures [44].

This chapter also describes the creation of a new highly challenging EL evaluation dataset which was mentioned briefly as a contribution in Section 1.3. This dataset is exemplary of the range of problems faced when applying NLP techniques to newly digitised CH resources. The content is extremely noisy, makes reference to numerous ambiguous or unimportant entities, and requires highly specialised domain knowledge to interpret. Other researchers may find this dataset interesting both as a specific example of a challenging CH corpus and a means of testing EL systems under relevant circumstances.

3.2 Evaluation Corpus

The corpus used for this evaluation is derived from a collection of 8,000 depositions gathered from the peoples of Ireland in the 17^{th} century in the aftermath of an uprising known as the 1641 Rebellion. The depositions document the various losses, military actions, attacks and transgressions inflicted on numerous individuals during the Rebellion. In spite of some controversy surrounding the accuracy of certain witness statements, the depositions provide a 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 [55] 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 spelling of words in the English language had not yet been normalised. 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

Three complete depositions in their original transcribed form have been added to Appendix A for interested readers.

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. [75] tackled the problem of normalising spelling in the depositions with great success, while the Cultura project [106] 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.

While Mitankin's work was extremely successful in normalising the language of the depositions, the choice was made to work with the original unnormalised text. The justification for this decision was that the use of the normalised text would set an unrealistic precedent. While we are fortunate that excellent tools exist for English, off-the-shelf normalisation software will not always be available e.g. in the case of medieval Latin. It was the opinion of the researcher that focusing on the unnormalised text would yield a more general solution to EL in this problem space.

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 presents 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.

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[1]. From the perspective of

[1] `http://dbpedia.org/page/Felipe_VI_of_Spain`

a naïve disambiguation tool, the modern king can seem like a much better referent than the monarch who is actually described in the depositions, namely Philip IV[2]. Koutraki et al. specifically study this problem, although the task of disambiguating the mention is still under investigation [62]. It is also worth noting 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 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 KBs and the under-representation of the entities that are relevant to the collection.

3.2.1 Corpus Preparation

From the complete collection of depositions 16 documents were sampled for use in the evaluation. Documents were chosen to be approximately 800 words in length as it was 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.

To help with this, some basic pre-processing was manually performed that, in practice, would be expected of an appropriate software library. Content that was tagged as being in the margins of a deposition or that had been crossed out by the original scribe was removed from the text. These were marked by `<note>` and `<del>` tags in the original TEI files making it easy to identify and remove these features. The depositions were also tokenised into approximate sentences as, again, this is an operation that could be automated given a suitably implemented tokeniser.

Using WebAnno [12], 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. When using GERBIL, any label in the gold standard corpus that cannot be found in GERBIL's KB is treated as a NIL annotation. Thus there is no required convention for marking a NIL annotation in the gold standard. However, the samples on the GERBIL wiki[3] use the following convention:

[2] `http://dbpedia.org/page/Philip_IV_of_Spain`
[3] `https://github.com/dice-group/gerbil/wiki/URI-matching#consequences`

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

where `entity_text` was the surface form of the entity with spaces removed. This convention was adopted in the annotation of the 1641 depositions. The annotations were checked for correctness by a historian who was an expert in the history of the rebellion.

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.

3.3 Experimental Setup

The evaluation of available EL systems was performed using GERBIL. GERBIL is an online evaluation platform which 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.

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. However, as the annotation services are hosted outside of GERBIL and the maintenance of these services is left to their respective hosts there is no guarantee that all services listed will be available when performing an experiment. Indeed, at the present time, some services (e.g. AIDA) have been fully deprecated and are no longer available for experimentation even though GERBIL lists them on the experiment configuration page.

The experiment configuration interface also allows users to upload custom datasets in NLP Interchange Format (NIF) [49]. WebAnno, which was used for annotating the depositions, does not support this format, hence the corpus was first exported as in XML Metadata Interchange (XMI) format and a parser was written to convert it to NIF. This NIF dataset was uploaded to GERBIL.

The experiment type chosen for this investigation was Disambiguate to Knowledge Base (D2KB). 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. The decision to run the experiment in this manner is due to the focus of this research being on 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.

However, it was found that even when configured to perform a D2KB task, some EL systems will still

perform NER. GERBIL accounts for this by applying a filter to the EL service's output and removing any entity mentions that do not match a mention in the gold standard. However, this still creates a problem when running an experiment as EL systems which attempt to perform NER will very likely suffer a penalty in their performance scores. This issue is described in greater detail in Section 3.5.

The depositions described in Section 3.2 were uploaded as a custom dataset to GERBIL and the experiment was configured to evaluate all available annotation systems against the collection using a D2KB experimental setup. Under these conditions GERBIL ran the experiment. GERBIL performs four types of evaluation: Emerging Entities, InKB, GSInKB and what will be termed a "standard" evaluation. These four evaluation types are described in greater detail in Section 3.4.

3.4 Defining Precision, Recall and F1 in GERBIL

Precision, Recall and F1 are, of course, well known measures of performance used to assess the quality of solutions in domains such as IR. While the definitions of these metrics in GERBIL are not different from their conventional form, it is important to understand how GERBIL computes them in order to appreciate the meaning of the results it presents. These definitions can be somewhat unintuitive for those coming from a different field of research. The typical definition of Precision, Recall and F1 measure do not change. That is to say:

$$P = \frac{tp}{tp + fp}$$

$$R = \frac{tp}{tp + fp} \tag{3.1}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

However, the manner in which true positives, false negatives and false positives are defined can be a somewhat counter intuitive.

Suppose GERBIL is tasked with performing a D2KB experiment. The EL services to be tested will be provided with the source text of the document and annotations which indicate the position of entities within the text. Each system being tested does not need to perform NER. It simply applies an annotation to each mention (either a URI or NIL) and returns these to GERBIL.

Consider the sample input/output presented in Table 3.1. The mentions in the second column have been sent to an EL service which has returned the corresponding URI or NIL annotations in the fourth column.

Mention ID	Mention	Gold Standard	Annotator Response
1	County of Wickloe	`dbp:County_Wicklow`	NIL
2	Colonell Toole	NIL	`dbp:Lawrence_M._O'Toole`
3	Lord of Ormond	NIL	NIL
4	Citty of Dublin	`dbp:Dublin`	NIL
5	pope of Rome	`dbp:Pope_Urban_VIII`	`dbp:The_Pope`
6	Ireland	`dbp:Ireland`	`dbp:Ireland`

Table 3.1: Sample EL system response and corresponding gold standard labels

The URIs in the third column represent the correct annotations according to the gold standard. How will GERBIL score this system under the four evaluations it performs?

3.4.1 Standard Evaluation

The standard evaluation considers all annotations in the gold standard and the EL service's response. The value of true positive will be a count of how many URIs or NIL labels in the dataset were correctly applied. For the sample in Table 3.1, mentions 3 and 6 have been correctly annotated, yielding $tp = 2$.

Strangely, in GERBIL under a D2KB experiment the values of false positive and false negative increase together for the standard evaluation. In this instance, mentions 1, 2, 4 and 5 have incorrectly been labelled giving $fp = fn = 4$

Ultimately this results in the following Precision and Recall values:

$$P = \frac{2}{2+4} = 0.333$$

$$R = \frac{2}{2+4} = 0.333 \tag{3.2}$$

$$F1 = \frac{2 \times 0.333 \times 0.333}{0.333 + 0.333} = 0.333$$

This, of course, leads one to ask, "when will false positive and false negative be different?" The answer is that these values will differ under a separate experiment type e.g. an A2KB experiment where the EL services must perform NER in addition to disambiguation. Under this experiment type, if an annotator fails to identify that a string corresponds to a mention then this will increase the value of false negative.

If the annotator erroneously marks a string of text as an entity then this will increase the value of false positive.

Under the standard D2KB experiment this situation can never arise as the annotator has already been told which strings to annotate. Hence the annotator cannot fail to spot a mention, nor can it accidentally mark irrelevant text as an entity. It can, however, apply an incorrect label to a mention resulting in the aforementioned simultaneous update in false positive and false negative. The consequence of this approach is that the values for Precision, Recall and F1 in the standard D2KB evaluation will always be the same, unless something has gone wrong (discussed in Section 3.5).

3.4.2 Emerging Entities

The Emerging Entities (EE) evaluation considers whether or not NIL labels have been appropriately applied to the mentions in the dataset. For the sample in Table 3.1, mentions 1, 2, 3 and 4 are considered in the Emerging Entities evaluation as either the gold standard, the EL system response or both are NIL for each of those mentions.

In this case, mention 3 is a true positive as the system has correctly applied a NIL annotation to the mention. Mentions 1 and 4 are both false negatives as the system applied a NIL label when it should have supplied a referent URI. Mention 2 is a false positive as the system incorrectly applied a referent URI when it should have supplied a NIL label. This results in $tp = 1$, $fp = 1$, and $fn = 2$, giving the following values for Precision, Recall and F1:

$$
\begin{aligned}
P &= \frac{1}{1+1} = 0.5 \\
R &= \frac{1}{1+2} = 0.333 \\
F1 &= \frac{2 \times 0.5 \times 0.333}{0.5 + 0.333} = 0.4
\end{aligned}
\tag{3.3}
$$

Note how the separate definitions of false positive and false negative here allow Precision and Recall to differ.

3.4.3 In Knowledge Base

Conversely to the Emerging Entities evaluation, the In Knowledge Base (InKB) evaluation tests whether or not referent URIs were appropriately applied to the input corpus. The mentions which are included in this test are mentions 1, 2, 4, 5 and 6 as either the gold standard, the EL system response or both have a referent for each of those mentions.

Mention 6 is a true positive for this evaluation as the system applied the correct URI to the corresponding mention, $tp = 1$.

Mentions 4 and 5 are false negative annotations. In the case of mention 4, this is because the system applied a NIL label when it should have supplied a referent URI. In the case of mention 5 the supplied referent URI does not match the URI in the gold standard. This gives $fn = 2$.

Mention 2 is a false positive as the EL system erroneously supplied a referent URI when it should have applied a NIL label. Thus $fp = 1$.

Putting these together yields the following for Precision, Recall and F1:

$$P = \frac{1}{1+1} = 0.5$$

$$R = \frac{1}{1+2} = 0.333 \tag{3.4}$$

$$F1 = \frac{2 \times 0.5 \times 0.333}{0.5 + 0.333} = 0.4$$

Again, note how the separate conditions for false positive and false negative allow Precision and Recall to differ.

3.4.4 Gold Standard In Knowledge Base

Finally, the Gold Standard In Knowledge Base (GSInKB) evaluation only considers the mentions that are not NIL in the gold standard dataset. Essentially it measures how many of the referents in the gold standard were found and correctly applied by the EL service. For Table 3.1 the mentions considered in this evaluation are 1, 4, 5 and 6.

The only true positive here is mention 6, thus $tp = 1$.

Similarly to the standard evaluation, mentions 1, 4, and 5 simultaneously increase the false positive and false negative counts giving $fp = fn = 3$. Hence the values for Precision, Recall and F1 are all the same in this evaluation:

$$P = \frac{1}{1+3} = 0.25$$

$$R = \frac{1}{1+3} = 0.25 \tag{3.5}$$

$$F1 = \frac{2 \times 0.25 \times 0.25}{0.25 + 0.25} = 0.25$$

In the context of an A2KB experiment, a false negative would occur if the EL system failed to spot that a string corresponded to an entity. Given the nature of the D2KB experiment, this scenario can never arise which leads to this equality in the GSInKB results.

3.4.5 Micro and Macro Precision, Recall and F1

GERBIL uses two different methods of computing precision and recall for EL problems, namely macro and micro Precision and Recall. These metrics have already been discussed in Section 2.9.1, but a brief review is given below.

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}.

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 and the values of P_{macro}, R_{macro} and $F1_{macro}$ are the average of P, R and $F1$ scores obtained.

In the event of a division by zero, 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[4] and is enforced by the use of the GERBIL tool for evaluating EL systems.

While this may seem inappropriate at first glance, it is a reasonably sensible design decision if one considers the case of an input document with no entities. The most appropriate thing for an EL system to do would be to annotate nothing. This would result in a division by zero error, but the EL system has done exactly what it should do and therefore should be awarded a score of 1. If the EL system *does* annotate a mention, then it should be awarded a score of 0 as it has incorrectly found entities.

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

3.5 Managing Disobedient Entity Linking Systems

While GERBIL can be configured to perform a D2KB experiment, this does not mean that every EL system included in the experiment will adhere to this experimental setup. Once a document and the marked entity mentions have been sent to an EL system, the target system is free to ignore the input mentions and may choose to perform NER, despite the fact that this defeats the purpose of the D2KB experiment.

GERBIL accounts for this by applying a strong boundary matching filter on the mentions returned from the EL systems. Consider, for example, the text below with the underlined text denoting the entity mentions as marked in the gold standard:

Thomas Parnell of the Cittie of Dublin gouldsmyth sworne and examined deposeth and sayth...

Assume a disobedient EL system performs NER and returns annotations for the mentions "Thomas Parnell" and "Dublin". The "Thomas Parnell" response is included in the evaluation as it exactly matches the mention in the gold standard. The "Dublin" mention is disregarded and counted as a false positive on the part of the EL system. This will happen even if the EL system identified the correct referent for the partially matched mention.

A simple litmus test to check whether or not a system has adhered to the D2KB problem is to look at its Precision, Recall and F1 values for the standard and GSInKB evaluations. If these values are not equal, then the system attempted to perform NER instead of adhering to the experimental setup.

At the time that this experiment was conducted, 8 of the available annotators were online [40, 112, 99, 78, 115, 105, 13, 69]. Of these only 2 were found to actually perform the experiment correctly. AGDISTIS (see Section 2.8.2 for description) and PBOH (see Section 2.8.1 for description) accepted the list mentions in the depositions and attempted to annotate them with URIs from DBpedia. All other EL systems attempted to perform NER and suffered severe performance penalties as a result.

While GERBIL does account for some level of disobedience on the part of the EL systems being tested, considering the incredibly noisy nature of the depositions, the margin for error here is too large. The results from the remaining 6 EL systems could not be considered fairly in this evaluation. Ongoing work is continuing to investigate the performance of the approaches implemented by these disregarded systems. For the interested reader, the original table of results are presented in Appendix B.

Given that this feature of GERBIL proved to be both problematic and difficult to detect, the maintainers of GERBIL were contacted with the suggestion of providing a more explicit warning where EL systems may have incorrectly carried out an experiment task. The maintainers agree that such a feature would be beneficial, but at the time of writing this request has not yet been implemented.

3.6 GERBIL Results and Discussion

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**
PBOH	0.4250	0.4250	0.4250	0.4266	0.4266	0.4266

Table 3.2: Results of standard evaluation in D2KB experiment obtained from GERBIL

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
PBOH	0.2696	0.2203	**0.3834**	0.2799	0.2292	**0.3594**

Table 3.3: Results of InKB evaluation in D2KB experiment obtained from GERBIL

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**
PBOH	0.5565	**0.7561**	0.4612	0.5782	**0.7542**	0.4688

Table 3.4: Results of Emerging Entities evaluation in D2KB experiment obtained from GERBIL

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
PBOH	**0.3834**	**0.3834**	**0.3834**	**0.3594**	**0.3594**	**0.3594**

Table 3.5: Results of GSInKB evaluation in D2KB experiment obtained from GERBIL.

Upon completion of an experiment, GERBIL returns a vast array of statistics for each of the four evaluation types. These have been organised and presented across Tables 3.2, B.2, B.3, and B.4. The decision was made to report on the Precision, Recall and F1 values for each experiment type for the purposes of demonstrating that the figures were gathered in a reliable manner. As mentioned in Section 3.5, equal values for P, R, and F1 can be used as a litmus test to determine whether or not an EL system carried out the experiment task correctly.

Due to the problems outlined in Section 3.5, this is not a comprehensive comparison of EL systems. However, given that much of the information that is available for informing an EL system is likely to

be built on relationships (see Chapter 4), for the purposes of this thesis the evaluation was sufficiently informative for progressing an invesigation.

Looking at the results in Table B.1, it would seem that AGDISTIS performed significantly better than PBOH. In particular it achieved the best performance in the EE task but actually performed quite poorly in the InKB task, albeit still a better performance than PBOH. This suggests that AGDISTIS' stronger performance in the standard evaluation was 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 results in Table B.3 where it achieved a macro and micro F1 score of 0.7189 and 0.7326 respectively.

While AGDISTIS exhibits the best performance on three of the tasks, it is outperformed by PBOH in the GSInKB task. This suggests that, while AGDISTIS is better at abstaining from annotating, PBOH is better at choosing the correct referent when such a referent exists.

Although neither annotator's scores are particularly impressive, AGDISTIS' results were deemed interesting enough to warrant further investigation. Its ability to abstain from annotating was considered highly desirable and in this capacity it greatly outshone PBOH. While it was clearly less accurate on the GSInKB task, the difference between PBOH and AGDISTIS was significantly narrower than on other tasks.

3.7 Detailed AGDISTIS Linking Process

This section will provide a detailed analysis of how AGDISTIS works in order to identify why it performed as well as it did in the evaluation task.

3.7.1 Indexing Process: Building the Knowledge Base

At the time that AGDISTIS was originally designed, state of the art EL systems typically assumed that Wikipedia or some Wikipedia based derivative would be used as the KB. This led to some assumptions on the part of the EL system about the information that would be available in the KB. AGDISTIS was intended to be a KB agnostic approach to EL. The KB indexes triples from a semantic knowledge base, but makes very few assumptions about the structure of the KB itself.

AGDSISTIS uses Lucene to index triples and later to retrieve triples during the linking process. Triples are stored across four fields in the index: subject, predicate, object_uri (for triples that map entities to entities), and object_literal (for triples that map a subject to a literal value such as a surface form). At run-time, AGDISTIS queries the KB using surface forms and URIs (executed as String queries over the search index). The KB returns relevant triples using standard IR relevance measures. The structure of the KB index, visualised using Luke[5], is shown in Figure 4.

[5] `https://github.com/DmitryKey/luke`

Additional surface forms for entities in the KB may be supplied to AGDISTIS during the the KB construction process in the form of a TSV file which maps URIs to surface forms. These surface forms are associated with the entity in the index using the `http://www.w3.org/2000/01/rdf-schema#altLabel` property as a predicate.

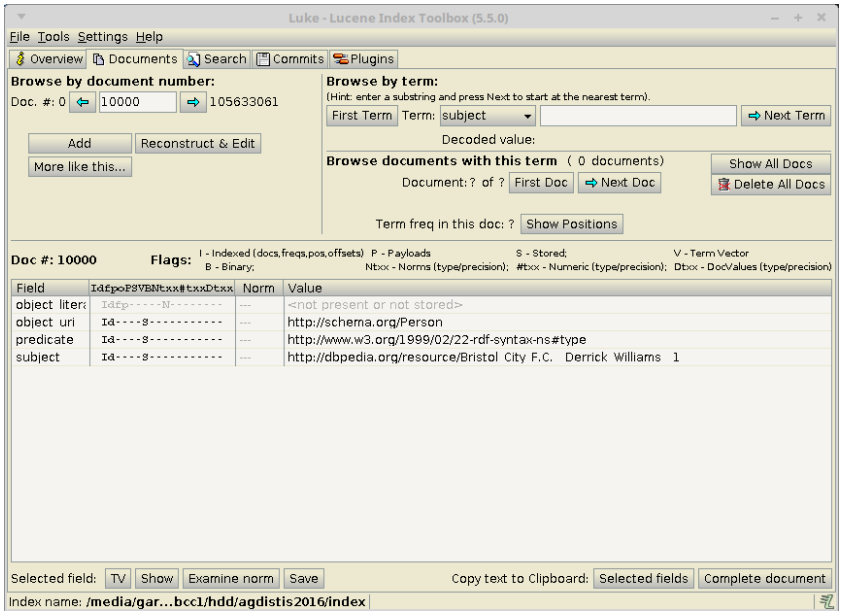


Figure 4: Structure of AGDSISTIS KB index

Optionally, AGDISTIS can also use a context based KB to search for candidate referents. The context KB stores the URI of the candidate referent, surface forms for the candidate referent and long form texts in which the candidate referent is mentioned. In other words, the index stores text that describes the context in which a mention occurs. The index also contains a field called URI Count, which is the number of contexts stored for a given candidate referent. As mentioned previously, use of the context index is optional. Even when enabled, this index is only used if a candidate referent could not be found in the default KB index. The structure of the context KB index, visualised using Luke, is shown in Figure 5.

The recommended DBpedia datasets for setting up a local deployment of AGDISTIS are:

- `disambiguations`
- `instance_types`
- `labels`
- `mappingbased_literals`
- `mappingbased_objects`

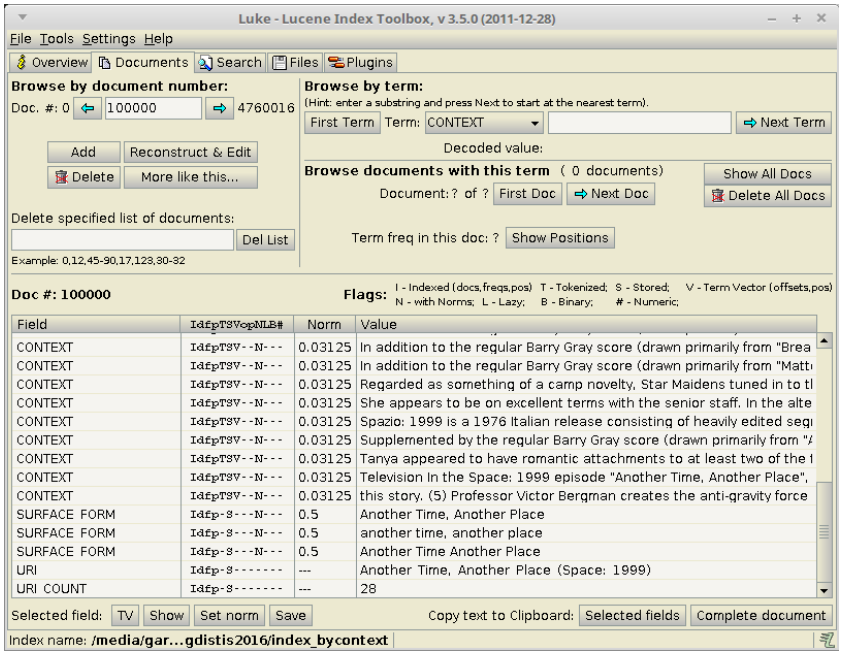


Figure 5: Structure of AGDSISTIS KB context index

- `persondata`
- `specific_mappingbased_properties`
- `transitive_redirects`

Details about each of these datasets can be found on their respective download pages on DBpedia. For the purposes of this thesis, the English version of each dataset was used in local deployments of AGDISTIS. Datasets were obtained from the DBpedia website[6].

3.7.2 Candidate Selection

The candidate selection process for AGDISTIS is highly configurable and is performed in a series of stages. Key to its operation is the fact that it will terminate its search for candidates as early as possible. As soon as one of its approaches to identifying candidates succeeds, its search will cease.

If AGDISTIS has been configured to search for candidates using acronyms, the candidate search begins here. If a mention is deemed to be an acronym (a short sequence of upper case letters), then the candidate selection process will attempt to identify entities in the KB which may correspond to that acronym. If the acronym search finds a set of candidates, then the process terminates here. Otherwise, if search by

[6] `https://wiki.dbpedia.org/downloads-2016-10`

acronym is disabled, or the search by acronym finds no candidates then the search proceeds to the next stage.

The initial search for candidates based on surface form similarity will only search for exact string matches. Mentions may be treated to strip trailing "s" characters or other similar lemmas, but otherwise AGDISTIS looks for entities with surface forms which exactly match the format of the entity mention. If any set of triples are found then they may conditionally be added to the graph (conditions are described below) and the search will terminate.

If the candidate selection process has not found any entities in the default KB index and AGDISTIS has been configured to use the context KB, then the surface forms are executed as queries against the context index. All mentions are concatenated and executed as a single query against the context index. The results returned are used to construct possible URIs that may be found in the default index. The default index is searched using these URIs and results are compared for similarity with the original mention. The threshold similarity for this check is extremely low at 0.3 and is hardcoded into AGDISTIS.

If, at this stage, the process still has not found any candidates then the process repeats, with a fuzzy threshold for the comparison between mentions and surface forms. In the first pass, only surface forms that are annotated with the predicate `http://www.w3.org/2000/01/rdf-schema#label` are considered as surface forms and they must match exactly. On the second pass, surface forms annotated with the predicate `http://www.w3.org/2000/01/rdf-schema#altLabel` are also considered.

The manner in which a label is checked for similarity with the entity mention is dependent on the predicate of the triple in the KB. Surface forms that have the predicate `http://www.w3.org/2000/01/rdf-schema#label` must be exact matches to the mention. Surface forms which have the predicate `http://www.w3.org/2000/01/rdf-schema#altLabel` need only be partial matches to the mention. Similarity is computed using Grzegorz Kondrak's N-Gram Distance measure [61] as implemented in Apache Lucene. The algorithm is set to use trigrams by default, but this can be configured. The threshold similarity for `http://www.w3.org/2000/01/rdf-schema#altLabel` is 0.87, but this can be changed. Altering the threshold make AGDISTIS more forgiving of spelling errors in mentions. Its default value is chosen by the developers of AGDISTIS who arrived at it empirically.

When a set of candidates are found, a series of checks are performed before they are added to the entity graph. Redirects are resolved to the entities that they redirect to, and disambiguation pages are ignored. Although AGDISTIS is KB agnostic, it does hardcode two predicates which it uses to achieve this. These are the predicates for DBpedia redirects and disambiguation pages (namely `dbp:wikiPageRedirects` and `dbp:wikiPageDisambiguates` respectively). If a candidate referent is found to be a redirect to another entity in the KB, then AGDISTIS will use the URI of the entity that it redirects to, rather than the URI retrieved from the KB. If a candidate referent is an instance of a Wikipedia disambiguation page, then the candidate is discarded.

It is possible to filter candidates based on their type in the KB as well. By default, AGDISTIS only considers entities that belong to a small pool of types including Person and Place. It was found that disabling this filter improved the quality of AGDISTIS' results on standard evaluation datasets such as AIDA-CoNLL [52], MSNBC [18] etc. This is largely due to the prevalence of a wide variety of entity types in the content of these collections. Limiting AGDISTIS to only a small subset of entity types results in it failing to identify otherwise good referents for a mention. In the DBpedia Spotlight corpus [69], word senses are annotated as well as named entities. Disabling the filter enabled AGDISTIS to identify referents for word senses, thus improving performance.

3.7.3 Referent Selection

After the pool of candidates has been selected, AGDISTIS begins to grow a graph-based on the relationships between entities. These relationships are obtained from the KB. HITS is then applied to the graph in order to identify the most suitable referent for each mention.

Growing the graph is achieved using Breadth First Search (BFS) with the candidate referents acting as the initial seeds for the graph. The search begins at the graph nodes corresponding to each candidate referent in turn. The KB is queried for all triples where the candidate referent's URI is the subject of a triple. The returned triples are filtered according to the prefix of the predicate. For example, the edge type might be defined as `http://dbpedia.org/ontology` meaning that only objects that are linked to the subject by predicates in the DBpedia ontology may be added to the graph.

The object of each filtered triple is added to the graph as a new node with a directed edge pointing from the subject to the object. Similarly to the edge type, the node type may also be filtered according to the prefix of the node's URI. By default this is `http://dbpedia.org/resource`

The maximum depth of BFS is configurable and is set at 2 by default.

Once the execution of BFS is completed, AGDISTIS runs HITS across the resulting graph. For each mention, the candidate referent with the highest authority score is chosen as the referent. Only the initial candidate referents can be chosen as referents for a mention i.e. none of the nodes added by BFS are considered as candidate referents for a mention. AGDISTIS aggregates the chosen referents into a map of results and assigns NIL to those mentions for whom a referent could not be found.

It is interesting to note that, aside from choosing the highest ranked candidate referent, AGDISTIS does not perform any filtering based on the hub or authority scores of the candidate referents. A mention which is assigned at least one candidate referent is guaranteed to return a chosen referent that is not NIL. So AGDISTIS' seemingly discerning ability to abstain from annotating where appropriate does not, in fact, stem from its referent selection process. Rather it is the fact that the candidate selection process prevents poor candidates from ever being added to the graph that helps AGDISTIS to achieve appreciable scores in the Emerging Entities task. This is illustrated in Figure 6 where a graph can be seen containing

candidates for only 4 out of 21 mentions in a deposition. The remaining 17 mentions are simply assigned a NIL label.

The graph in Figure 6 was generated by downloading and running a local instance of AGDISTIS. The source code of the project was modified to add a function which ran just before the referent selection process returned its chosen set of candidates. This captured the state of the graph upon completion of AGDISTIS' referent selection process and exported the state of the graph to a file in GEXF format.

3.8 Examining Candidate Retrieval in AGDISTIS

As was noted in Section 2.9.2, a problem with the evaluation metrics used by GERBIL is that they conflate sources of error which can make it difficult to determine why a given EL system performed poorly in a task. In the case of AGDISTIS, the rendering of the disambiguation graph in Figure 6 highlights two things:

1. Many candidates have no links, or only have outbound links
2. Most of the mentions sent to the service have no candidates

The first of these points is an issue with the KB and a lack of available information. This is not surprising given the problems with entity coverage in the KB which have already been discussed. This issue is addressed in Chapter 4.

The second point, however, is interesting because it suggests that AGDISTIS' success in the Emerging Entities task is due to its candidate selection process being brutally selective of candidates, rather than because its referent selection process is good at dealing with ambiguity.

Hachey's evaluation metrics presented in Section 2.9.3 provide a measure of the performance of the candidate selection process in terms of Precision and Recall. These measures can be used to investigate the impact of AGDISTIS' candidate selection process on its overall EL performance.

It must be emphasised that in this context, Hachey's Meaures are being employed to shed light on the reason behind AGDISTIS' success, rather than to determine the quality of AGDISTIS' linking method.

3.8.1 Measures

Recall Hachey's evaluation metrics as described in Section 2.9.3:

$$\mu_C = \frac{\sum_i |C_i|}{|M|} \tag{3.6}$$

Node Label	Entity Mention	URI
1	Richard Phillips	dbp:Richard_Philipps
2	Richard Phillips	dbp:Sir_Richard_Phillips
3	Richard Phillips	dbp:Richard_Phillips_(chemist)
4	John Dickenson	dbp:John_Dickerson_(trainer)
5	John Dickenson	dbp:John_Dickenson_(author)
6	Richard Phillips	dbp:Richard_Phillips_(merchant_mariner)
7	Richard Phillips	dbp:Richard_Phillips_(MP)
8	Ballylynan	dbp:Ballylinan
9	Richard Phillips	dbp:Richard_Phillips_(English_painter)
10	whitewall	dbp:Whitewall_tire
11	Richard Phillips	dbp:Richard_Phillips_(American_painter)
12	John Dickenson	dbp:John_Dickenson_(Canadian_politician)
13	Richard Phillips	dbp:Richard_Phillips_(athlete)

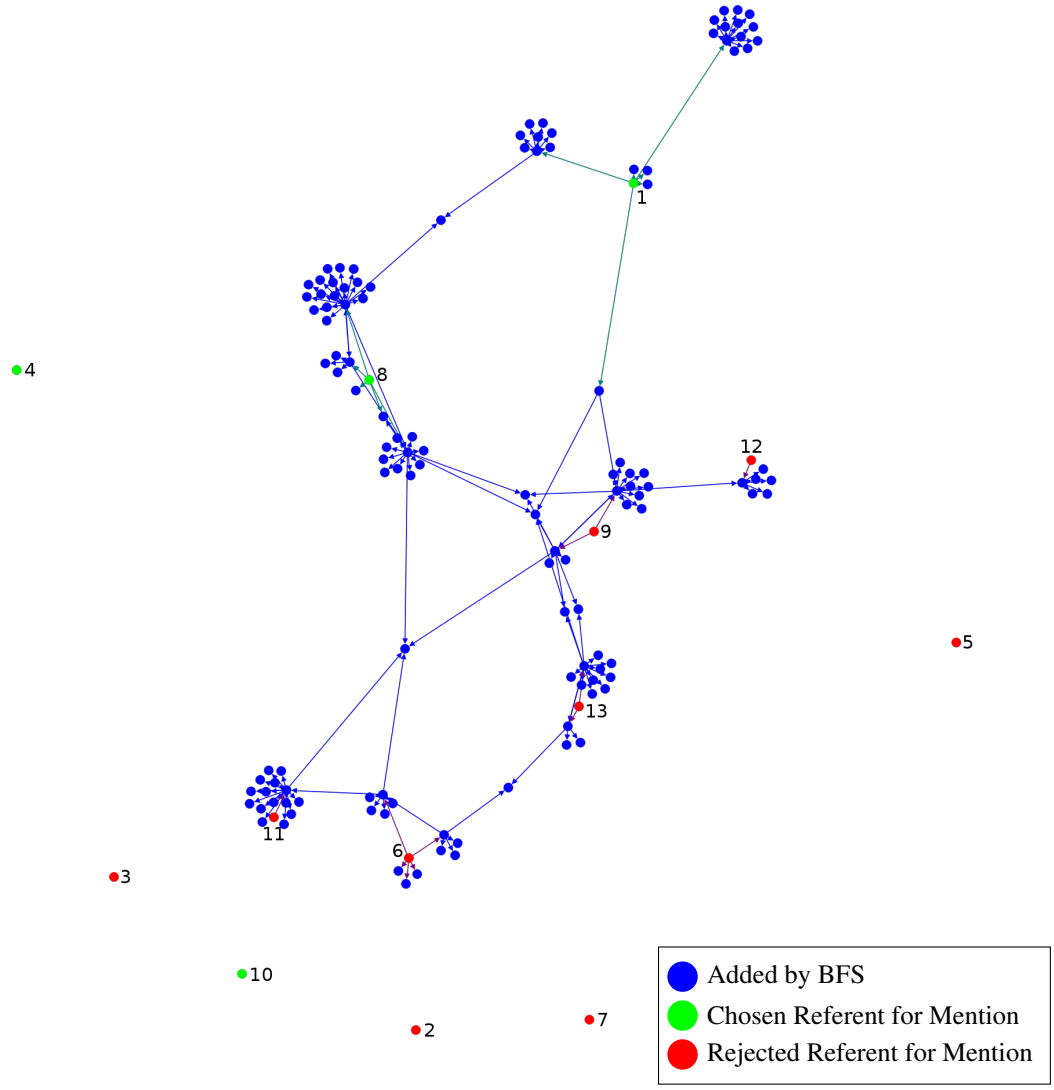


Figure 6: Visualisation of graph created and analysed by AGDISTIS when performing entity linking on Deposition 812296r239. Note that there were 21 unique mentions in the text being annotated. The 17 entities not accounted for in this diagram received a NIL label.

$$P_C = \frac{|\{C_i \,|\, C_i \neq \varnothing \wedge \mathcal{G}_i \in C_i\}|}{|\{C_i \,|\, C_i \neq \varnothing\}|} \tag{3.7}$$

$$R_C = \frac{|\{C_i \,|\, \mathcal{G}_i \neq \text{NIL} \wedge \mathcal{G}_i \in C_i\}|}{|\{\mathcal{G}_i \,|\, \mathcal{G}_i \neq \text{NIL}\}|} \tag{3.8}$$

$$P_{\text{NIL}} = \frac{|\{C_i \,|\, C_i = \varnothing \wedge \mathcal{G} = \text{NIL}\}|}{|\{C_i \,|\, C_i = \varnothing\}|} \tag{3.9}$$

$$R_{\text{NIL}} = \frac{|\{C_i \,|\, \mathcal{G}_i = \text{NIL} \wedge C_i = \varnothing\}|}{|\{\mathcal{G}_i \,|\, \mathcal{G}_i = \text{NIL}\}|} \tag{3.10}$$

These metrics focus on the ability of the candidate selection process to identify relevant candidates in the KB, rather than on the ability of the EL system to return the correct referent for each mention.

The value of μ_C indicates how much ambiguity on average the system has to deal with for each mention that is to be processed. This is found by computing the mean size of each candidate set. For analysing AGDISTIS, a second definition of μ_C was used which did not include empty candidate sets when computing the average. This was due to the fact that mentions which have an empty candidate set do not ultimately influence the outcome of the referent selection process. It was a matter of interest to see how large the candidate sets that ultimately formed the disambiguation graph grew. The alternative definition is given as:

$$\mu_{C \text{ not NIL}} = \frac{\sum_i \{|\, C_i \,||\, C_i \neq \varnothing\}}{|\{C_i \,|\, C_i \neq \varnothing\}|} \tag{3.11}$$

Additionally, Hachey's definition of Precision and Recall were extended to include macro and micro Precision and Recall using the same definitions as GERBIL.

$P_{C\,micro}$, $R_{C\,micro}$, $P_{\text{NIL}\,micro}$ and $R_{\text{NIL}\,micro}$ were computed by concatenating AGDISTIS' output for every document in the evaluation corpus into a single file and evaluating the candidate selection process for the entire collection.

$P_{C\,macro}$, $R_{C\,macro}$, $P_{\text{NIL}\,macro}$ and $R_{\text{NIL}\,macro}$ were found by calculating P_C, R_C, P_{NIL} and R_{NIL} for each individual document in the evaluation corpus and then computing the average of these values

Hachey's metrics omit the F1 measure that usually follows Precision and Recall. Even so this value was computed as the harmonic mean of Precision and Recall yielding the four definitions below:

$$F1_{C\,micro} = \frac{2 \times P_{C\,micro} \times R_{C\,micro}}{P_{C\,micro} + R_{C\,micro}} \tag{3.12}$$

$$F1_{C\,macro} = \frac{2 \times P_{C\,macro} \times R_{C\,macro}}{P_{C\,macro} + R_{C\,macro}} \tag{3.13}$$

$$F1_{\text{NIL}\,micro} = \frac{2 \times P_{\text{NIL}\,micro} \times R_{\text{NIL}\,micro}}{P_{\text{NIL}\,micro} + R_{\text{NIL}\,micro}} \tag{3.14}$$

$$F1_{\text{NIL}\,macro} = \frac{2 \times P_{\text{NIL}\,macro} \times R_{\text{NIL}\,macro}}{P_{\text{NIL}\,macro} + R_{\text{NIL}\,macro}} \tag{3.15}$$

3.8.2 Evaluating Candidate Retrieval in AGDISTIS

A local deployment of AGDISTIS was modified to write the candidate set for each input mention to a TSV file on a per-request basis. The complete deposition evaluation corpus was executed against AGDISTIS and the results were aggregated in a series of TSV files.

A Python script was written to implement Hachey's measures according to the definitions given in Section 3.8.1. The output of the script after analysing the candidate sets output by AGDISTIS are given in Table 3.6.

	μ_C	$\mu_{C\text{ not NIL}}$	P_C	R_C	$F1_C$	P_{NIL}	R_{NIL}	$F1_{\text{NIL}}$
Micro	1.5292	5.3577	0.3796	0.2708	0.3161	0.6501	0.7743	0.7068
Macro	1.5550	8.1619	0.4147	0.2773	0.3324	0.6640	0.7742	0.7148

Table 3.6: Results for Hachey's measures as applied to AGDISTIS

The results of this test clearly highlight how AGDISTIS performs well in the Emerging Entities task. Its candidate selection process excels at blocking irrelevant entities from being considered as candidates as demonstrated by the high values for P_{NIL} R_{NIL} and $F1_{\text{NIL}}$. However, the extremely low scores for P_C, R_C and $F1_C$ in both the macro and micro evaluations shows that the process is overzealous and often omits the correct referent from the pool of candidates. Hence, irrespective of how good HITS is at choosing the correct referent, it cannot possibly return the appropriate URI because it is never added to the entity graph.

3.9 Summary

Clearly a large problem with annotating a collection such as the depositions is the lack of representation for the 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 KBs, much like Brando's approach [9] which built bridges between DBpedia, BnF and other sources of information. However this must be done with an eye to the Emerging Entities problem. Given the

ad-hoc nature of the entities encountered in the depositions (often servants or soldiers), identifying an all-encompassing knowledge base will be difficult if not impossible. For those entities who simply do not exist in the KB resources available, these mentions must remain unannotated. The challenge of identifying suitable sources of knowledge is tackled in the next chapter.

Regarding the performance of AGDISTIS, it is useful to note that the system is not necessarily effective because HITS is good at choosing a referent from an ambiguous pool of candidates. It is effective because the candidate selection process filters candidates aggressively, keeping the resulting graph of candidates and relationships reasonably small. This is advantageous both because it makes the selection process easier for HITS and keeps processing time down. However it is problematic because it is overzealous and frequently filters the correct referent from the pool of candidates. Furthermore HITS itself is overly optimistic. If even one candidate is chosen for a mention, then a referent for that mention will be returned. There is no way to filter candidates based on the authority or hub scores provided by HITS.

This chapter has also described the creation of a new EL evaluation corpus. This is an extremely challenging corpus which represents a number of the primary challenges faced when working with CH material. The references to entities are ambiguous, niche and often require an awareness of the time period, or highly specialised domain knowledge, in order to accurately assign a referent. The language is noisy which, in this case, is due to the developing nature of the English language, but in alternative corpora could be due to issues such as OCR errors. The corpus is available for use online[7] and future work is intended to expand it to include more sample depositions and annotations from a broader range of KB sources.

[7] https://github.com/munnellg/1641DepositionsCorpus

Chapter 4

Building a Knowledge Base

"The whole panoply of the universe has been neatly expressed to them as things to (a) mate with, (b) eat, (c) run away from, and (d) rocks."
— Terry Pratchett, *Equal Rites*

4.1 Overview

The manual annotation of the gold standard described in Chapter 3 demonstrated the severe lack of coverage of entities in DBpedia with respect to the 1641 depositions[1]. The survey of available KBs in Chapter 2 illustrates the diverse range of alternative KBs that exist, but there is no individual source that could be used to suitably annotate the depositions.

For highly specialised CH collections, this is not an atypical problem. Newly digitised records from previous centuries can be so highly specialised that information about the entities they contain is, as yet, emerging and will not be documented in established KB sources. However, this does not mean that information about these entities is not documented *somewhere* within the annals of historical scholarly outputs.

Even so, it is impossible to compile a compendium of all things that have ever existed. At some point, no matter how complete the KB, an entity mention will be encountered for which there is no suitable known referent. The appropriate action to take here from the perspective of an EL system is to apply a NIL label, to indicate that the entity is unknown. However, there is an interesting related challenge here which is Knowledege Base Population. Here, if a previously unidentified entity is encountered, a new entity may be created in the KB to denote that a reference to a previously undiscovered entity has been found. Generally this is not how EL works, and so this problem is not tackled in this thesis.

[1] A mere 23% of deposition entities had corresponding DBpedia entries

Rather, in order to tackle this issue of lack of entity coverage, research turned to the construction of new KB resources that might provide better coverage than those currently available. At the very least, a KB should diminish the number of erroneous annotations provided by the EL service by removing entities which are not relevant to the collection.

While there might be an expectation that a good EL system will cut through ambiguity and noise in the KB to arrive at an appropriate set of referents, the structure of the KB can have a drastic effect on the quality of entities returned during candidate selection. In the case of niche collections it can be advantageous to filter the contents of the KB such that only a core pool of entities relevant to the corpus are indexed and retrieved. If nothing else, this filtering gives the corpus curator a degree of control over the magnitude of expected errors in the system's output.

Underlying this investigation is an attempt to construct a KB that might be deemed reputable from the perspective of scholars who are considered experts in the source material of the collection. Wikipedia is a common point of contention which can be cited here as an example of an ill-reputed source. Irrespective of how accurate the content of Wikipedia articles has become through constant refinement and communal updates, the fact remains that Wikipedia is considered a dubious source of information from the perspective of scholarly research. It was the experience of this PhD that historical scholars were sceptical of any KB that was derived from Wikipedia.

Yet another requirement on a newly constructed KB is the need for it to integrate with other existing KBs that may be more established. If a collection is annotated with a highly specialised KB such that it cannot integrate with other collections, then half the purpose of the exercise has been lost. EL adds structure to a collection, but it also facilitates integration with other existing collections, assuming a relationship between the vocabulary of both collections can be established.

This chapter will focus on efforts to construct a new KB from resources used by historical scholars during their academic pursuits. Candidate resources for the construction of the KB were identified on two fronts. Primary sources, which were obtained directly from the time period being investigated, and secondary sources comprised of essays written by prominent experts in the history of Ireland and the United Kingdom. Ultimately the primary sources are deemed to be too noisy and too subject to debate for them to be useful as KBs. However two new KBs based on secondary sources are generated and later used for EL.

In the context of this work, a primary source is a source which is directly derived from an original CH item. For example, the depositions are a primary source even though they are in digitised TEI format because the text they contain is the original text extracted from the 17^{th} century documents. Similarly a database of banking transactions or census data would be considered a primary source provided it stored the exact, unmodified information that was present in the original document.

A secondary source is a source which does not come directly from the time period being studied. Rather

it is a source which was created at a later date that discusses, describes, or otherwise documents the events of the chosen time period. These would be books, essays, biographies etc. written by historians.

The process transitions content from secondary source biographies to a KB source for EL is not bespoke. In this chapter it is applied to two separate sources of information with little modification other than to extract content from the web pages of the respective sites and acknowledge minor formatting differences in the presentation of entity names. This process is one which could be followed, both in order to construct a new KB source and integrate this new source with other existing KB sources.

This chapter also documents the linking process developed during this PhD which facilitated the integration of the two new KBs with DBpedia. The linking method is described in Section 4.7.

4.2 Primary Sources

Given the focus on 17^{th} century Ireland and, in particular, the focus on the 1641 Irish rebellion, three primary sources relating to the events of the rebellion were identified and investigated for use as KBs with the aid of a historian who is an expert in the time-period. These were The Statute Staple, The Down Survey, and The Books of Survey and Distribution (BSD).

The main focus of these primary sources is on land ownership, making them a useful source of information about geographic regions, but less helpful when determining the identities of 17^{th} century people. The Down Survey provides a list of important societal figures in the form of prominent landowners, but it is certainly not a comprehensive list of Ireland's inhabitants.

The challenge faced with respect to identifying individuals is tackled more directly through secondary sources (discussed in Section 4.5), but the stark reality is that identifying people is orders of magnitude more challenging than identifying locations. Perhaps through a long, iterative process it may be possible to create a KB which documents all recorded peoples of Ireland across the centuries, but a more tractable solution is to simply determine which people are worth documenting, and which are too ignominious to warrant discussion. Callous as this may sound, it may be the most practical solution from the perspective of the challenges faced in these collections.

A resource which was identified but ultimately not used was John Lodge's "The Peerage of Ireland, or a Genealogical History of the Present Nobility of that Kingdom" [66], which documents the genealogy of significant Irish families up to the 18^{th} century. While a list of names for individuals could be generated from the index of the peerage, reliably constructing a KB from its contents was determined to be a greater task than could be completed within the scope of this thesis. Nevertheless, this resource should be earmarked and considered for future work.

4.2.1 Down Survey

The Down Survey is a survey of land ownership in Ireland conducted by William Petty between the years 1656–1658. The purpose of the survey was to accurately establish land holdings by prominent Catholics and Protestants in Ireland. This was done in order to determine how Catholic land would be redistributed in recompense for their involvement in the 1641 rebellion.

The Down Survey has been the subject of much previous research and historians have generated lists of landowners which they believe uniquely identify all landowners recorded in the Down Survey. Each landowner has been assigned a unique identifier, making the Down Survey a useful source of entities which are known (or at least believed) to be distinct.

The landowners list is split across two years: 1641 and 1670. This is because the Down Survey represents land holdings both before and after the 1641 rebellion. From the perspective of historical scholarship the two lists are not compatible as they contain different lists of names due to various landowners perishing or having their holdings confiscated. However, from a technical standpoint, the IDs assigned to each name are generally consistent. That is to say, the same ID maps to the same name between the two sets with some occasional spelling variations e.g. ID 227 maps to "Alderman Kennedy" in 1641, and to "Alderman Walter Kennedy, of Finnstown" in 1670.

However, there is no certainty that the same name refers to the same person across the two time spans, given that a father's name often passed to their heirs. It must therefore be assumed that there is no equivalence between the entities in the two lists and that they have not been disambiguated. This is a practical constraint that was recommended by the historians who generated the data.

Consequently when performing linking with respect to the Down Survey, the time period of the source must be considered and linking must be performed with respect to the appropriate Down Survey landowners list. The recommended threshold year for choosing between the two lists is 1652, as by this time most Catholic merchants had been expelled from Ireland and their lands seized under "The Act for Settling Ireland" [104].

The Down Survey documents the ownership of slightly less than 60,000 distinct properties in Ireland[2] and identifies 6,676 distinct landowners in 1641 and 3,853 distinct landowners in 1670. Taking the intersection of the 1641 and 1670 landowners lists (optimistically assuming that an ID always refers to the same person across the two sets of records) gives a total of 9,093 unique individuals mentioned in the Down Survey.

Not all landowners recorded in the Down Survey are people. In some instances an organisation such as "School of Armagh" is listed as the owner of a property. Some properties are deemed to own themselves e.g. "Unprofitable Bog" is listed as a landowner in both 1641 and 1670 with ID 308. Somewhat more problematic is the inclusion of phrases such as "All of Above" which is listed as a landowner with ID

[2] 59,585 land records to provide the exact number.

Property	Quantity
Holdings	59,585
1641 Landowners	6,676
1670 Landowners	3,853
Total Landowners	9,093

Table 4.1: Summary of statistics for Down Survey

6834. The final example is even given a religious denomination ("Protestant"). Hence there is some noise in the Down Survey, but it is a useful starting point for the construction of a KB.

County	Holdings	County	Holdings	County	Holdings	County	Holdings
Antrim	1,731	Fermanagh	2,158	Mayo	3,183	Wexford	2,299
Armagh	970	Galway	4,056	Meath	1,583	Wicklow	1,354
Carlow	561	Kerry	2,672	Monaghan	1,847	Unspecified	111
Cavan	1,977	Kildare	1,208	Offaly	1,143		
Clare	2,221	Kilkenny	1,573	Roscommon	2,042		
Cork	5,325	Laois	1,111	Sligo	1,285		
Derry	1,226	Leitrim	1,483	Tipperary	3,153		
Donegal	2,637	Limerick	1,919	Tyrone	2,144		
Down	1,281	Longford	872	Waterford	1,569		
Dublin	882	Louth	675	Westmeath	1,334		

Table 4.2: Breakdown of documented land holdings in the Down Survey by County

In addition to landowner IDs, historians working with the Down Survey have also assigned unique identifiers to properties listed in the survey. As noted in Chapter 2 a problematic aspect of modern geographic KBs is that they often omit the six Northern Irish counties, grouping them with the United Kingdom. The Down Survey, of course, makes no such distinction, and all 32 Irish counties are present. A breakdown of the number of records per county is given in Table 4.2. The IDs used are obtained from OSI and OSNI [3], meaning that the Down Survey is compatible with official modern sources of information about locations in Ireland.

Analysing the Down Survey gives an interesting indication of where modern geographic KB may be limited with respect to Ireland. Given that the Down Survey breaks Ireland down into roughly 60,000 distinct regions, the completeness of other geographic databases can be approximated by comparing how

[3] Note that OSNI has since been amalgamated with the Land and Property Services in Northern Ireland

many distinct locations they present. A caveat here is that the regions in the Down Survey are extremely granular, sometimes denoting specific fields, bog-lands, and small properties.

Considering Geohive, GeoNames and DBpedia which were mentioned in Chapter 2, the following values can be derived.

Geohive is comprised of 50,607 townlands, parishies, baronies and counties demonstrating that it has excellent coverage of Ireland. Again, it should be noted that this only includes the Republic of Ireland as Northern Ireland is handled by a separate mapping agency who have not released their data in a semantic format.

For DBpedia, the number of locations in Ireland was approximated by downloading the DBpedia Geo Coordinates dataset from the 2016 DBpedia dataset and loading the data into a triplestore. For the purposes of this research, Parliament v2.5[4] was chosen as the triplestore. Parliament was selected because it supports GeoSPARQL queries over triples, which was an important consideration given that the data is geographic in nature. GeoSPARQL enables querying over triples based on their relative spatial positions. Thus a GeoSPARQL query was executed which searched for points that fell within the bounding box illustrated in Figure 7[5], identifying which points in the DBpedia dataset were locations on the island of Ireland and the surrounding islands.

This resulted in the identification of 7,600 locations in Ireland that are represented in the DBpedia Geo Coordinates dataset.

GeoNames has 26,457 location names in its IE dataset for the Republic of Ireland. Filtering the UK dataset so that it contained only those points that fell within the previously defined bounding box yielded 1,607 locations for Northern Ireland for a total of 28,064 locations on the island of Ireland.

Hence it can be seen through the analysis of a primary source record where modern semantic sources of information about geographic locations in Ireland may be limited.

The consequence of these limitations is that a large proportion of locations identified in cultural heritage sources cannot be annotated with a URI from GeoNames or DBpedia. Effectively these locations must be given a NIL label unless an alternative KB can be identified which documents their existance.

Of course, much like people, not all locations are of equal importance. Major regions of the country i.e. counties are, of course captured. However there is a poor representation of smaller geographic units such as townlands, and parishes. The impact of the absence of these locations from GeoNames and DBpedia depends on how frequently they are mentioned throughout a given corpus.

[4] `https://github.com/SemWebCentral/parliament`

[5] Coordinates are: (lng: -8.569, lat: 55.392), (lng: -5.988, lat: 55.392), (lng: -5.317, lat: 54.782), (lng: -5.317, lat: 54.150), (lng: -6.152, lat: 52.107), (lng: -10.031, lat: 51.200), (lng: -10.898, lat: 52.160), (lng: -10.283, lat: 54.451), (lng: -8.569, lat: 55.392)

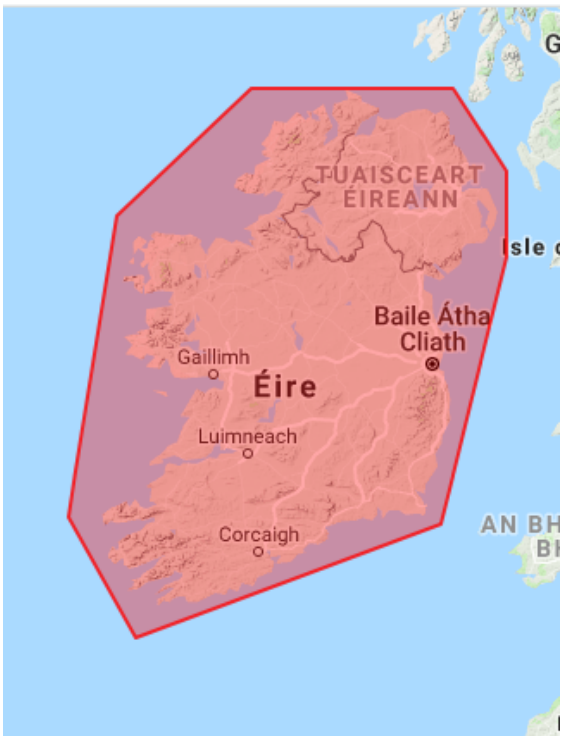


Figure 7: The highlighted region shows the area in which points were sought from the DBpedia Geo Coordinates dataset and the GeoNames dataset.

4.2.2 Statute Staple

The Statute Staple documents financial transactions between the years 1597–1687. Several prominent landowners and their respective debts are mentioned in the ledger. From a historical perspective, the Staple is interesting as it suggests motives for Irish landowners to engage in rebellion. From the perspective of EL it provides an indication of relationships between landowners during the 17^{th} century. As discussed in Chapter 2, relationships can be a useful feature for EL systems.

Unlike the Down Survey, entries for individual people and locations in the Statute Staple have not been assigned unique identifiers, making it difficult to determine exactly who was involved in a given transaction. The Staple documents 3,993 transactions, where each transaction can involve multiple debtors or creditors. Not everyone in the Staple is a landowner, however. Hence there will be a number of debtors and creditors whose identity cannot be discerned from the Down Survey.

Even so, the Staple can be used to determine relationships between the landowners mentioned in the Down Survey, provided an entry in the Staple can be assigned an ID from one of the two master lists of Down Survey landowners. Given the time span of the Staple and the previous threshold given in Section 4.2.1, transactions before 1652 are linked with respect to the 1641 master list, and those during and after 1652 are linked with respect to the 1670 master list.

4.2.3 Books of Survey and Distribution

Similar to the Down Survey, the Books of Survey and Distribution (BSD) are records of land ownership around Ireland in the 17^{th} century. These records were used for taxation purposes, determining how much money an individual owed to the British government based on the number and size of properties they held.

The digital copies of BSD are not as well organised as those of the Down Survey. Unique IDs have not been assigned to land owners or locations, making it difficult to establish how much useful information the survey contains. An approximate quantification for the number of properties documented per county can be reached as illustrated in Table 4.3. However, these values are known to be erroneous and are larger than the true value of documented holdings due to duplicate entries in BSD. For example, the property Ballynaleney (Down Survey ID 2991) in Co. Antrim occurs twice in the BSD records for Antrim. In truth, even if duplicate entries were not a problem, it is still unlikely that the number of records in BSD would match those in the Down Survey.

Working with BSD serves to highlight one of the problems encountered when working with raw primary source material. History is noisy, and primary source materials will conflict with each other. Computer scientists can propose broad compromises which may seem like solutions, but these solutions can be naïve from the perspective of historical research. The input of historical scholars at this level is crucial to ensuring the correctness of any outputs produced.

The approximate figures presented in Table 4.3 were obtained by manually partitioning the records in BSD into individual spreadsheets for each county. This task was performed by a historian, Dr. David Brown, whose primary research interests involve landownership and taxation in 17^th century Ireland. Dr. Brown has previously been involved with the Petty Maps project and aided both in the mapping of Down Survey records to Ordnance Survey Ireland records, and the identification of landowners for the landowners lists mentioned in Section 4.2.1.

The values reported below are the number of entries in each spreadsheet per county. This includes duplicates. The result is a total of 64,302 identified regions on the island of Ireland.

While it is meaningful to link the Down Survey to BSD for the purposes of historical research, in the context of KB construction, it was found that much of the information found in BSD such as land holdings and relationships such as co-ownership had previously been established by the Down Survey. BSD could be used to confirm or contest the historical accuracy of the Down Survey, but it was not helpful for EL.

4.3 Resolving Entities Across Primary Sources

Given the three primary sources described in Section 4.2, the challenge is to resolve instances of entities across the sources in order to extract useful information for a KB. Effectively, this is a record linking problem [27, 100, 25]. Records across Down Survey, Statute Staple and BSD can be compared for

County	Holdings	County	Holdings	County	Holdings	County	Holdings
Antrim	2,315	Fermanagh	2,377	Mayo	5,269	Wexford	1,420
Armagh	1,048	Galway	7,122	Meath	2,090	Wicklow	631
Carlow	748	Kerry	2,121	Monaghan	2,427		
Cavan	1,766	Kildare	1,410	Offaly	1,074		
Clare	6,025	Kilkenny	1,178	Roscommon	3,904		
Cork	3,439	Laois	943	Sligo	1,553		
Derry	977	Leitrim	1,110	Tipperary	2,674		
Donegal	880	Limerick	1,735	Tyrone	1,312		
Downe	1,576	Longford	1,049	Waterford	764		
Dublin	1,324	Louth	574	Westmeath	1,467		

Table 4.3: Approximate breakdown of documented land holdings in the Books of Survey and Distribution

equivalence based on the names of landowners and the names of locations. The methods by which these comparisons were performed are described below in Sections 4.3.1 and 4.3.2.

Instances of people and places in BSD were mapped respectively to landowner IDs and Location IDs from the Down Survey with the comparison being performed based on the landowner name and location name respectively. In the case of people, historians had manually tagged whether an individual mentioned in BSD owned the property in 1641 or 1670. The landowner ID was chosen from the appropriate Down Survey master list according to these tags.

Work on BSD proceeded in stages, with the initial objective being to link the names of locations in BSD to location IDs in Down Survey. After this process was complete, the names of landowners in BSD were mapped to the names of landowners in Down Survey. Note that the names for the owners of properties did not always agree between the two resources. In these instances, a mapping was not created and it was the task of the historian to determine the most appropriate course of action based on best practice in their field of study.

Statute Staple was linked to Down Survey based on the names of Debtors and Creditors. The choice between linking with respect to the 1641 master list or the 1670 master list was determined by the date of the transaction. All records prior to 1652 were linked to the 1641 list and all records after and including 1652 were linked to the 1670 list.

For each comparison, a threshold similarity was empirically chosen to determine whether or not a link would be established. The metric chosen for this similarity measurement was Monge-Elkan distance, whose implementation details are described below in Section 4.3.1. All matches falling under this similarity threshold were rejected outright. Where multiple record resolutions had the same similarity, all

candidates were listed in the output and a manual selection process was performed. Indeed, all generated mappings were subject to a manual review by a historian in order to ensure the accuracy of the output.

This manual review relied on the expertise of the historian. They examined the mappings directly, made a judgement call based on their own best practices. In the case of mappings with only one candidate, if the two names of the individuals were close enough in spelling (where "close enough" was a subjective measure based on what the historian felst was appropriate) then the mapping was accepted. If there were multiple candidate mappings, then it was the duty of the historian to use their own judgement to determine which mapping should be accepted.

Usually this simply involved checking to see if any of the entities had an address (e.g. if the BSD listed an individual as residing in Meath, and of the two options, only one landowner was from that same county, then it was clear which was the correct mapping). Otherwise the mapping may be left blank, or connected based on some other factors known to the historian.

4.3.1 Resolving Landowner Names

Resolving the names of landowners between sources is challenging, with inconsistencies in spelling conventions and use of various titles presenting the greatest problem. A method of comparing fields which accounted for these variations was required. While initial tests were run using only Jaro-Winkler similarity, this approach was found to be too pessimistic and failed to capture some ostensibly good matches e.g. "Alderman Walter Kennedy, of Finnstown" should match "Alderman Kennedy". The solution was to use the Monge-Elkan Method [76] in combination with some preprocessing of the name strings.

Initially, each name underwent a simple normalisation process to eliminate some of the variations in how titles were written. Some surnames for example separate common prefixes from the rest of the name e.g. "Fitz Patrick" instead of "FitzPatrick". In such cases the space between the two parts of the surname were removed. Hyphens were added between the components of titles e.g. "brigadier general" became "brigadier-general". These transformations were achieved using a series of regular expressions.

For each name, information such as titles, honorifics and locations were variously collapsed and removed from the name string using a series of regular expressions and a gazetteer. Crucially, intermediate stages of the collapsing process were also stored and used for comparison. This generates a set of permutations which may be alternate surface forms for the person in the record. For example, a name such as "Alderman Walter Kennedy, of Finnstown" is reduced to "Walter Kennedy", but a comparison is also performed on the intermediate forms "Alderman Walter Kennedy", "Walter Kennedy, of Finnstown", as well as the original string. The scripts used to generate these permutations are available on Github along with the gazetteers which listed normalisation transformations, honorifics and titles[6].

The similarity between two strings S_1 and S_2 is computed using the Monge-Elkan method [76].

[6] `https://github.com/munnellg/RecordMatcher`

Choosing Monge-Elkan makes sense when one considers the nature of the problem at hand. Consider the examples "John FitzPatrick" and "Lord Jon FitzPatricke", two names which are ostensibly alike. It is possible that both mentions are references to the same person and it would be helpful if a string comparison function could capture their apparent similarity.

Levenshtein [65] and Jaro-Winkler [117] are inappropriate as they consider the strings in their totality. Both similarity measures will be confounded by, for example, the presence of "Lord" at the start of the second string. It would be more beneficial if the similarity measure could operate on smaller units of the string i.e. the different parts of the name. Set comparison measures such as Dice [23] or Jaccard [57] may seem like good candidates, but they will only work if the strings that make up each part of the name are exact matches. What is needed is a weighted, set comparison metric to determine if two sets have similar contents, rather than the exact same elements.

What is being described, broadly, is Monge-Elkan.

To implement Monge-Elkan, the name strings are split on whitespace yielding two sets of tokens $\mathcal{T}_1$ and $\mathcal{T}_2$. The sets are added to a bipartite graph with edge weights computed using Jaro-Winkler similarity [117]. An optimal mapping $\mathcal{T}_1 \mapsto \mathcal{T}_2$ is found using Edmond's blossom algorithm [26] giving $\mathcal{W}$, the set of weighted edges which comprise the mapping. Name similarity (denoted below as Φ) is the generalised mean of the edge weights in $\mathcal{W}$ as described by Jimenez et al. [59] where $m = 2$ in this experiment:

$$\Phi(S_1, S_2) = \left(\frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} w^m \right)^{\frac{1}{m}} \tag{4.1}$$

For any comparison of two records, all permutations of the name string are compared and the pairing which yields the highest score is returned as the final similarity of the two records. This will, of course, increase the number of false positive matches in the dataset, but this was an expected outcome. Because the resulting matches were subject to manual review, the focus was on increasing recall over precision.

4.3.2 Resolving Placenames

Resolving the names of locations was considerably less challenging than that of matching landowner names. The kinds of spelling variations, use of honorifics, titles etc. that were seen with people are not a problem when dealing with the names of places. Comparison of placenames was thus conducted using the Monge-Elkan method as described previously, without the additional effort of producing permutations of the string. The script for performing this task is available on Github[7].

[7] `https://github.com/munnellg/RecordMatcher`

4.4 The Problem With Primary Sources

In spite of the fact that much work has been done on resolving entities between the resources discussed in Section 4.2, results are not yet ready for use as EL KBs. It must be understood that one of the objectives of the work in this thesis is to annotate collections in a manner that is trustworthy and reliable for scholars. When annotating an entity with a URI, it is not just a consistent identifier that is added to the content of the text. It is all the attached information that comes with linked data. If the facts in the KB source are highly questionable, then it is not appropriate to use it as a KB. Consider the frustration of an academic who is constantly misled by the annotations supplied by an EL service.

While the record resolution described has made it possible to build bridges between these collections, and a certain amount of high level information that may be useful for an EL system could be extracted, a KB built on this work would, at present be too noisy and subject to much debate. Work to improve the standard of linking within the resources is still ongoing. But the need to rigorously check all information as it is linked means this process is extremely slow and manual even with the assistance of record linking methods.

4.5 Secondary Sources

Given the challenges surrounding using primary sources, two highly reputed secondary sources were investigated for use as a KB. These were the Dictionary of Irish Biography (DIB) and the Oxford Dictionary of National Biography (ODNB), which are repositories of historical biographies relating to the British Isles.

Both DIB and ODNB are collections of biographies written by historians about notable Irish and English historical figures. The subject of each article is usually a single entity which corresponds to a person. Titles contain the subject's forename, surname and variant names, and links between related biographies exist in the text of each article. Hence they exhibit structural properties similar to those that originally made Wikipedia a useful KB for EL. They are of greater specificity to the history of the British Isles than other more general resources and thus may help to fill some of the gaps that exist in KBs such as DBpedia. At the very least they may limit the scope of the linker's search to entities that are relevant to this geographic region and time period?

As previously mentioned, the biographies are written by notable historical scholars, making them a well reputed source among researchers. While any historical source is open to some amount of debate, the fact that these are highly reputed secondary sources means that the information they contain is somewhat less disputable than that of the primary source material previously encountered.

4.6 Extracting Information from Biographies

Both The Dictionary of Irish Biography (DIB) and The Oxford Dictionary of National Biography (ODNB) are available to browse and read online via their respective web portals. In order to obtain copies of the biographies, a crawler was written which iterated over the indexes of the sites and scraped the content of the listed biographies. Each biography was downloaded as an individual HTML file and automatically examined to extract useful information about the individuals they described. After scraping the biographies it was found that there are 9,642 biographies in DIB and 74,512 biographies in ODNB.

While the biographies in DIB and ODNB are not strictly structured, there are a number of consistencies in their formatting which make it possible to extract 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 from DIB:

"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 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. An illustration of these permutations is given in Figure 8.

Similarly, the date of birth for each individual is 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

Daniel Joseph (‘Dan’) Bradley

Daniel Bradley

Daniel Joseph Bradley

Dan Bradley

Dan

Figure 8: List of surface forms generated given the input “Bradley, Daniel Joseph (‘Dan’).

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”. Accounting for these variations is done by implementing an appropriate regular expression once the anomalies were detected. An approximate distribution of biographies in DIB and ODNB by century is given in Figure 9 and Figure 10 according to the extracted birthdays.

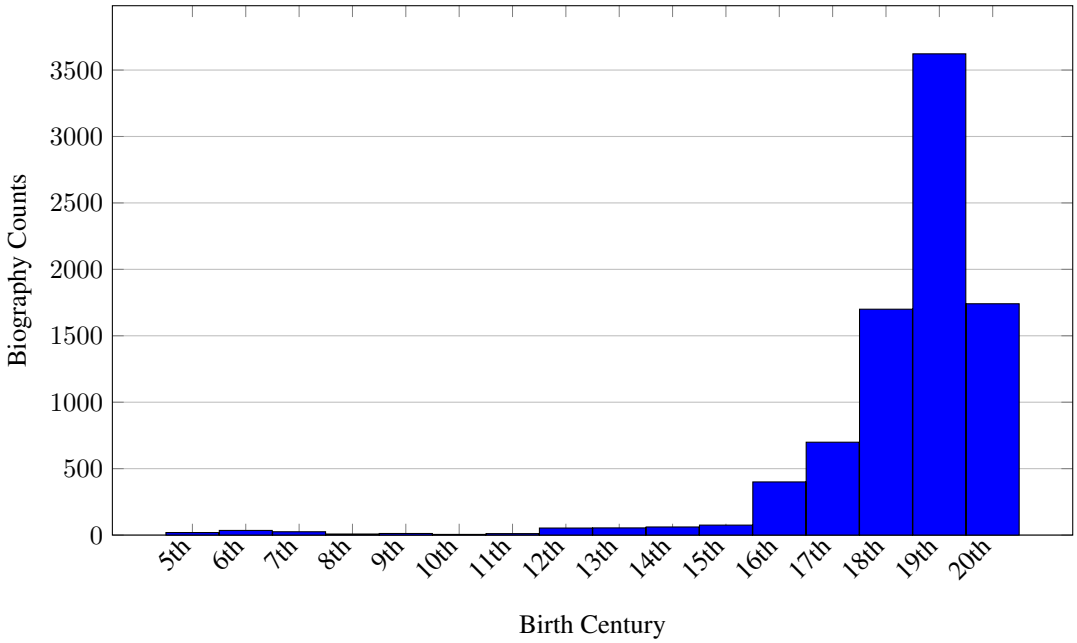


Figure 9: Distribution of biographies by century in DIB

Relationships between entities are often an important feature employed by EL systems when discerning the referent for an entity mention. Hyperlinks between biographies 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 vo-

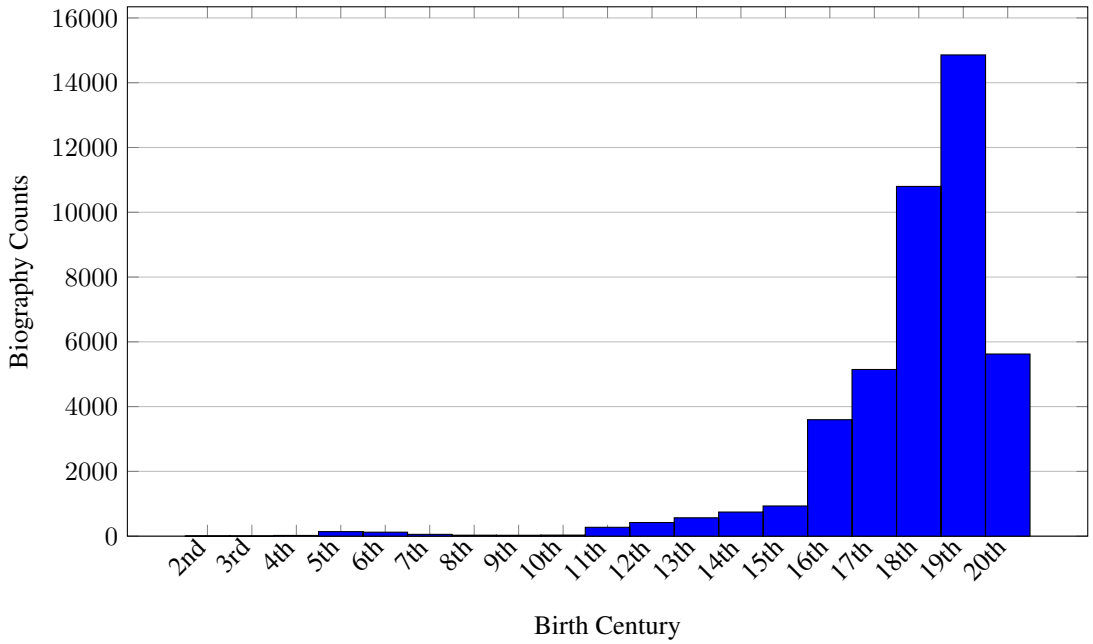


Figure 10: Distribution of biographies by century in ODNB

cabulary 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 and ODNB. 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 [14] 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 11 and a screenshot of the biography from which the information was extracted is in Figure 12.

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 these were not used but are certainly worth exploring in future work.

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

:time1248-01-01_1248-12-31 a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1248-01-01"^^xsd:date ;
    crm:P82b_end_of_the_end "1248-12-31"^^xsd:date .

:time1223-01-01_1223-12-31 a crm:E52_Time-Span ;
    crm:P82a_begin_of_the_begin "1223-01-01"^^xsd:date ;
    crm:P82b_end_of_the_end "1223-12-31"^^xsd:date .

:eventsbirth_a1292 a crm:E67_Birth ;
    crm:P4_has_time-span :time1223-01-01_1223-12-31 ;
    crm:P98_brought_into_life :Theobald_Butler_a1292 .

:eventsdeath_a1292 a crm:E69_Death ;
    crm:P100_was_death_of :Theobald_Butler_a1292 ;
    crm:P4_has_time-span :time1248-01-01_1248-12-31 .

:Theobald_Butler_a1292 a dbo:Person,
        crm:E21_Person,
        foaf:Person ;
    rdfs:label "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 :Richard_de_Burgh_a1131,
        :Theobald_Butler_a1291,
        :Theobald_Butler_a1293 ;
    owl: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" .
```

Figure 11: Sample RDF in Turtle format derived from “Butler (le Botiller, Pincerna), Theobald (c. 1223-1248)”

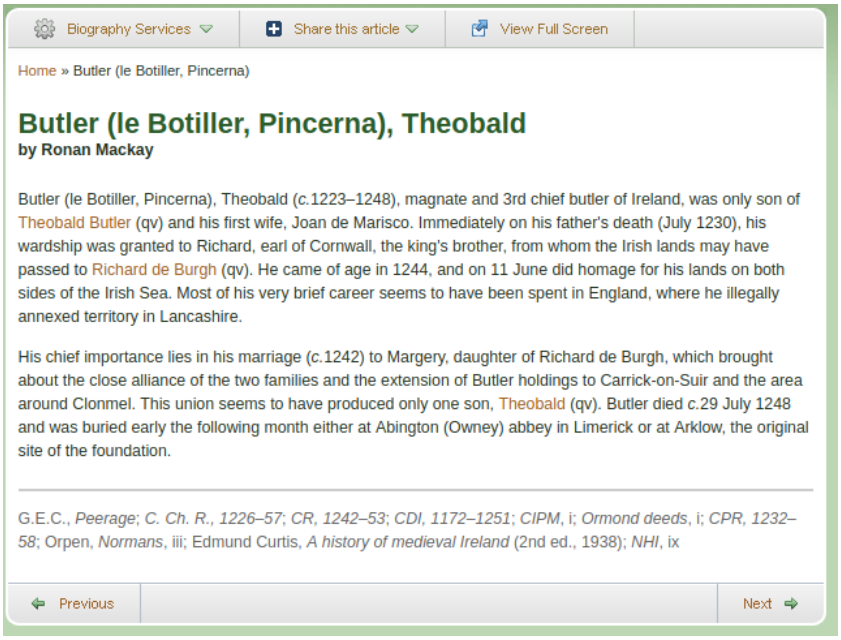


Figure 12: DIB Biography for Theobald Butler

4.7 Linking to Established Knowledge Bases

Recalling that one of the benefits of EL is that it enables linking between disparate sources provided they are built on the same vocabulary, a problem can be observed with using DIB and ODNB in their current format. In order to facilitate the integration of a KB derived from ODNB or DIB with DBpedia, an approach for linking biographies to their DBpedia counterparts was developed.

First, all DBpedia entities belonging to the class `dbo:Person` are indexed using Solr[8]. The name of each entity, the full text of the Wikipedia article from which they are derived, and anchor text on incoming links to the article were indexed.

Anchor text indicates alternative surface forms which may refer to an entity. For example, the DIB biography for the 7^{th} Earl of Mayo uses his full name and excludes his title, "Dermot Robert Wyndham-Bourke" while his name in DBpedia is given as "Dermot Bourke 7^{th} Earl of Mayo". Indexing anchor text can help to loosely capture the equivalence of these two references, assuming that Wikipedia uses the anchor text "Dermot Robert Wyndham-Bourke" to link to the Earl of Mayo's Wikipedia article from some other resource. However, it can also introduce some unwanted noise. For example, the anchor text for "Mountrath" has been found to point to the entity "Sir Charles Coote". Using anchor text as a source of surface forms can thus be something of a double-edged sword and it is worth investigating whether or not the effects of indexing this information are ultimately beneficial for a specific use case. In the context

[8] `http://lucene.apache.org/solr/`

of this work, it was found to be helpful in the identification of additional candidate referents.

For each biography entry in ODNB and DIB $b \in \mathcal{B}$, the title b_{title} is executed as a query against Solr. Matches on the title field and anchor text are boosted over matches in the article's content. A list of up to ten top-ranked candidates $\mathcal{P}_b$ is returned. The best matching DBpedia referent $p_b^* \in \mathcal{P}_b$ for a given biography is the one that maximises the expression:

$$p_b^* = \underset{p \in \mathcal{P}}{\operatorname{argmax}}\, \Psi(b, p) \tag{4.2}$$

Where $\Psi(b, p)$ is computed as a linear combination of content similarity and name similarity.

For a given candidate $p \in \mathcal{P}_b$, content similarity Ω between the biography $b_{content}$ and the candidate's Wikipedia article $p_{article}$ is computed using negative Word Mover's Distance (WMD) [63] as implemented in gensim [94]. This method establishes a vector representation of documents using word embeddings and then computes the distance between points in the two representations. Essentially, the dissimilarity of two documents is measured by examining how far the vector representations of words in one document must travel through space before the document will semantically match its counterpart. This is obviously a very computationally expensive operation. Similarity is found by subtracting the normalised distance from 1. Word embeddings are computed using a Word2Vec model [71] trained on a Wikipedia dump excluding redirects, disambiguation pages etc.

As with the comparison of names between primary source records, the name similarity function Φ is based on the Monge-Elkan Method [76]. The biography title b_{title} and name of a candidate p_{name} are lower-cased and tokenised. Stop words are removed yielding two sets of tokens $\mathcal{T}_b$ and $\mathcal{T}_p$. The sets of tokens are compared for similarity as described previously with $m = 2$ as before.

$$\Phi(b, p) = \left(\frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} w^m \right)^{\frac{1}{m}} \tag{4.3}$$

This yields the final formulation of Ψ as a function of the form:

$$\Psi(b, p) = \alpha\, \Phi(b_{title}, p_{name}) + \beta\, \Omega(b_{content}, p_{article}) \tag{4.4}$$

Where α and β are tuning parameters chosen such that $\alpha + \beta = 1$. Tuning these parameters essentially controls how much influence over the final result is exerted by the similarity between the surface forms of the two articles, versus the similarity of their respective contents. It is, of course, important to consider that there is a degree of similarity between surface forms, but the similarity between the content of a

Wikipedia article and a biography are what is really driving the computation of the similarity value. Configuring the values of α and β controls how much favour should be lent to either in the final linear combination.

A hard threshold τ is applied to p_b^*, enforcing a minimum similarity between a biography and its final chosen referent $\overline{p}_b^*$:

$$\overline{p}_b^* = \begin{cases} p_b^*, & \text{if } \Psi(b, p_b^*) > \tau \\ NIL, & \text{otherwise} \end{cases} \tag{4.5}$$

NIL indicates that a biography does not have a DBpedia counterpart.

The use of Monge-Elkan in combination with WMD is a useful symbiotic relationship.

Monge-Elkan is a string similarity function that compares the spelling of two strings. It can be used to check if two strings are made up of words with similar spellings, but not if the two strings are semantically similar.

WMD compares the semantic similarity of two documents by representing them using word embeddings and computing the distance between the vector representations of each word the document contains. This is helpful for establishing if two texts are related by some underlying theme, but not helpful for determining if a mention is a good match for a surface form in the KB.

Hence, Monge-Elkan is used to find entities in the KB who are candidate referents for the name attached to the biography (i.e. people whose names were spelled similarly to the name of the person in the biography). WMD was used to compare the semantics of the content of the biography with the content of a Wikipedia article that described the entity to determine which Wikipedia article (with a set of candidate articles having been identified with ME) is most likely to be describing the same entity as the biography.

It has been stated repeatedly that an important property of an EL system is that it should abstain from annotating if no suitable referent is found among the pool of candidates. For this linking method, the hard threshold provides a concrete definition of what is deemed to be "good enough" in terms of a minimum similarity between a biography and its best candidate referent.

As can be seen in the evaluation in Section 4.8, it is possible to find an optimal threshold which yields a best overall performance on the content of the biographies by varying the threshold and examining the effect on the quality of annotations applied to a gold standard corpus.

4.8 Evaluation of Linking Method

The approach described is essentially an EL solution. The service receives as input a surface form and some context which may help to identify the subject of the reference. Solr performs the candidate selection process, identifying a subset of candidates to which the surface form might be referring. The linking method then proceeds to identify the most likely referent from the pool of candidates. This means that it is possible to evaluate the performance of the method using EL benchmarking tools.

The linking service as it is currently implemented does not have a RESTful API endpoint. Rather it was implemented to process each of the biographies on disk as JSON files. Consequentially it could not be tested with GERBIL. However, The original implementation of the BAT framework was able to test the linking method by examining CSV files that were output by the tool. Hence the evaluations in this section were performed using the BAT Framework[9] [16].

Two ground truth, gold standard subsets were derived from a random sample of 200 biographies obtained from both DIB and ODNB (400 samples in total). A human annotator manually linked each sample with a corresponding DBpedia URI if an equivalent entity could be identified in the DBpedia ontology. Where no URI could be established, a NIL label was applied.

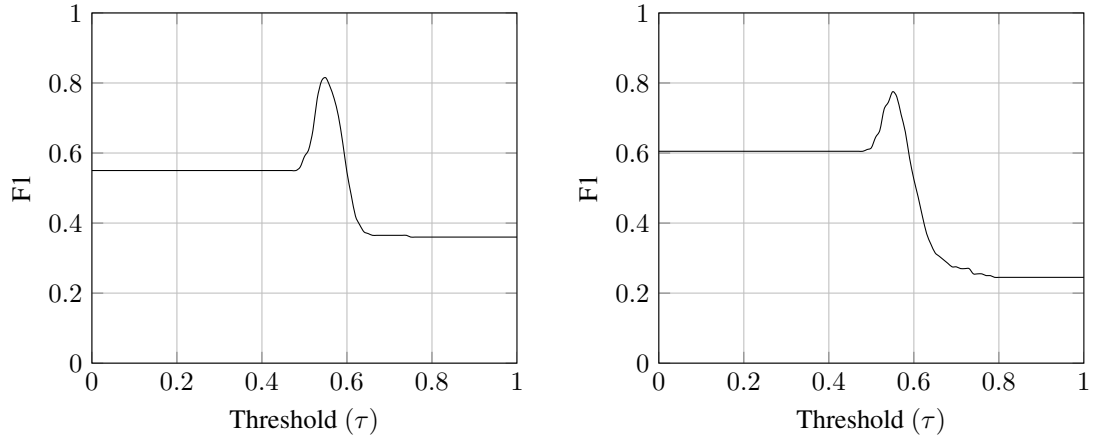
In order to match the samples with DBpedia, the annotator extracted the name of the individual described by a biography, dates of birth and death where provided and any key phrases which may help to identify them e.g. “Earl of Ormond”. These terms were searched for using Wikipedia and Google. Because DBpedia’s URI scheme matches that of Wikipedia, if an article for a person was identified in Wikipedia then the corresponding DBpedia entity could be identified. Sometimes a Wikipedia article was found without a matching DBpedia entry. In these cases the annotator marked the biography as not having an entry in DBpedia.

Identifying if a biography was a match for a DBpedia entity involved reading the content of both the biography and the Wikipedia article to ensure that both were describing the same individual.

Ultimately 64 of the ODNB samples and 72 of the DIB samples were labelled as NIL. This would suggest that approximately 36% of entities in DIB and 32% of entities in ODNB are not documented in DBpedia, indicating the extent to which ODNB and DIB can improve the scope of available knowledge to an EL system with respect to Irish people. Again, it should also be remembered that this KB has the effect of limiting the scope of the EL system’s search to a geographic region, which is undoubtedly beneficial.

For the purposes of the evaluation the values of α and β were fixed at $\alpha = 0.1$ and $\beta = 0.9$. This choice of weighting was due to the fact that a comparison with the name has already been partially performed by the candidate selection process. The strongest feature for identifying a referent is thus a comparison of the description of the entities as provided in the biography content and the text of the Wikipedia article.

[9] `https://github.com/marcocor/bat-framework`



(a) Change in performance on DIB with values of τ sampled at intervals of 0.01.

(b) Change in performance on ODNB with values of τ sampled at intervals of 0.01.

Figure 13: Performance of linking method on DIB and ODNB. Note that optimal threshold is seen when $\tau = 0.55$.

Even so, it was found that lending some small weight to the similarity between surface forms yielded a slight increase in the $F1$ score for the linking method. Experimenting with a number of different values eventually resulted in settling on those used for the experiment.

In the context of this experiment, the definition of F1 is the same as it was during the GERBIL experiment in Section 2.9.1 i.e. it is the harmonic mean of Precision and Recall. However, the fact that there is only one entity per document means that the results of both the micro and macro evaluations will be equal.

The method was tested by evaluating the quality of the links established by the method for varying values of τ. A threshold similarity of $\tau = 0.55$ was found to give the best results. This threshold yields the best trade-off between the method annotating a biography with a DBpedia URI or a NIL label. However, as can be seen in figures 13a and 13b, the method is highly sensitive to the value of τ, with a slight variation resulting in a dramatic drop-off in performance.

Arguably, given the need for accuracy when constructing KBs for academic study, a sub-optimal threshold $\tau > 0.55$ may be desirable. This will result in fewer overall links to DBpedia, but makes the algorithm more conservative, reducing the number of false positives.

It can be seen in Figures 13a and 13b that on either side of the optimal threshold, the performance of the system rapidly stabilises to a constant value. This is because the threshold only dictates whether or not the chosen referent should be returned. It has no effect on the actual choice of the referent. Hence, by setting the $\tau = 0$, for example, the linking tool will return all referents that it has identified, irrespective of how good or bad they are. For $\tau < 0.48$ or $\tau > 0.75$, this is essentially what has happened, resulting in the plot levelling off.

During the initial evaluation, subject to the conditions above, this approach achieved an $F1$ score of 0.815 on DIB, but only 0.675 on ODNB. Some of the imprecision stems from Solr as 43.1% of incorrect labels on ODNB and 45.9% of incorrect labels on DIB can be ascribed to the correct referent not being among the results returned by the search engine. However the remaining disparity in performance was somewhat alarming and subject to investigation.

It was found that the problem arose from multiple articles in ODNB which do not contain text. They are simply pictorial renderings of their subject. Consequently, the WMD algorithm had no content by which to compare the biography to Wikipedia articles. A follow-up investigation generated a new gold standard subset for ODNB with a minimum threshold of 50 words on the content of the biography for inclusion. The performance of the method improved dramatically on this collection, but still lagged slightly behind that of DIB with an $F1$ score of 0.775. The remaining disparity was ascribed to two challenging article types in ODNB:

1. ODNB contains disambiguation pages which list individuals who have the same surname. Identifying these pages programatically is challenging and so it is difficult to filter them.

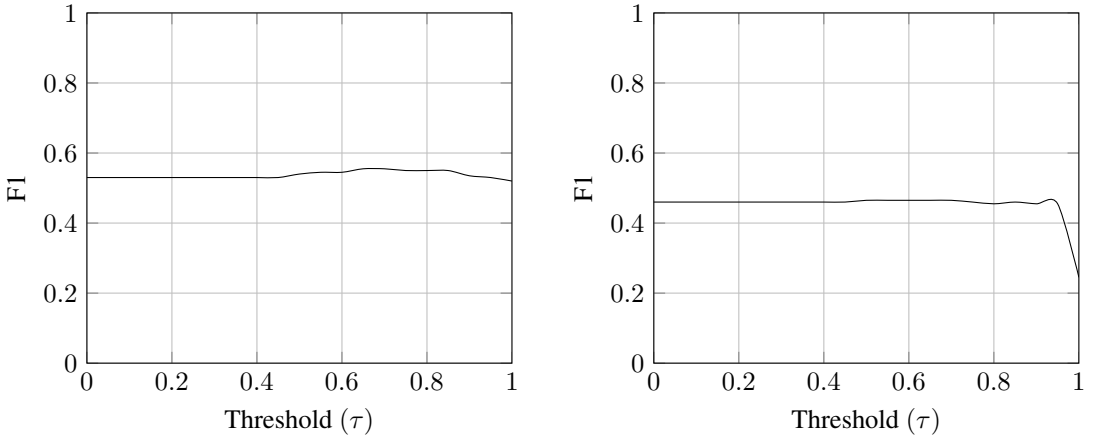
2. Some articles discuss more than one person, where multiple entities' stories are inextricably linked e.g. the famous serial killers Burke and Hare. Note that this is also a problem with DIB.

Collection	τ	F1
DIB	0.55	0.815
ODNB	0.55	0.675
ODNB (filtered)	0.55	0.775

Table 4.4: Summary of results

4.8.1 Further Analysis

In an attempt to evaluate the relative performance of this linking method with respect to other state of the art EL systems, a comparative analysis was conducted with respect to DBpedia Spotlight (see Section 2.8.3 for implementation details). The choice to use Spotlight was based on its application to a similar problem in the extraction of entities from Dutch biographies [35], and the fact that, of the EL systems that were known, it was the most likely to be able to tackle this specific problem. Its suitability for this task can be seen by noting that each biography is an EL problem that involves the resolution of a single entity surrounded by a large volume of text which describes the entity. This clearly gives an advantage to EL systems which perform analysis on the context of a mention rather than coherence between entities (e.g. the approach taken by Moro et al. [78]). As there is only one mention, the concept of coherence is somewhat meaningless.



(a) Change in performance of DBpedia Spotlight on DIB with values of τ sampled at intervals of 0.01.

(b) Change in performance of DBpedia Spotlight on ODNB with values of τ sampled at intervals of 0.01.

Figure 14: Results for linking by DBpedia. Note that the performance is much more consistent than earlier plots.

Due to the formatting of the entity mention at the start of each biography (leading surname followed by forename with various alternate names interspersed), it was noted that it may be difficult for DBpedia Spotlight to identify referents, given that the choice of candidates during the referent selection process involves looking for entities in the KB with similar spelling to the mention. The formatting of the mention would throw Spotlight off, resulting in an unfair comparison. In order to remedy this the surface form which named the subject of the biography was injected at the beginning of the article with some modification:

- The names were originally in surname, forename order. This was reversed.
- Where multiple formulations of a name were present between parenthesis e.g. different languages or nicknames, these alternate formulations were collapsed and removed from the surface form.

This was intended to yield a surface form which was more easily recognisable by Spotlight. For example, it may be difficult for Spotlight to identify a referent given the mention "Burke, Uilleag ('Uilleag an Fhiona') (Ulick de Burgh)". However, by removing all name parts that are between parenthesis and flipping the order of the names so that the forename comes first the mention "Uilleag Burke" is produced. This is much more manageable, and even a simple Google search will throw up the correct referent[10].

The evaluation was performed using Spotlight's disambiguation API endpoint[11] and used a custom script to submit the content of each biography individually along with the injected surface form as previously

[10] `https://www.google.com/search?q=Uilleag+Burke`
[11] `http://model.dbpedia-spotlight.org/en/disambiguate`

described. The responses from the server were dumped to a series of CSV files. As with the evaluation of this thesis' linking method the value of the confidence threshold for annotation was varied. Under these conditions, BAT reported F1 scores between 0.25 and 0.465 for the filtered ODNB dataset and scores between 0.52 and 0.555 for DIB depending on the value of the confidence threshold which ranged from 0 to 1. A summary of these results can be seen in Table 4.5 while Figures 14a and 14b show the effect of varying the threshold with each request. It is notable these graphs are much more stable than values shown in Figures 13a and 13b.

Collection	τ	F1
DIB	0.65	0.555
ODNB (filtered)	0.65	0.465

Table 4.5: Summary of results for DBpedia Spotlight

Given the task at hand, the method of linking presented by this thesis seems to identify referents in DBpedia with a reassuring level of accuracy. Indeed, the method is not restricted to this simple use case, as it is for all intents and purposes a fully implemented EL system. Given a set of surface forms and a context it should provide a set of suitable referents for the inputs.

However, this method falls into a common EL trap which is the trade-off between performance and time. The more accurate an EL method is, the more computationally expensive it is can become. This is true with this approach which requires as much as a minute to identify a referent for a single entity.

While this approach was initially conceived as an ad-hoc solution to a specific problem, its performance in the evaluation is encouraging and future work may seek to further investigate the construction of an EL service based on this approach provided the issue with time and computational complexity can be resolved. The current implementation is known to perform several wasteful operations, the results of which could be cached or even pre-computed and indexed to improve performance.

4.9 Summary

This chapter has focused on the acquisition and construction of different knowledge resources for CH focused EL. In the context of Irish history, there is a need to construct new semantic models that are targeted at specific aspects of Irish CH.

The two ontologies developed based on ODNB and DIB are available for download on Github[12]. Additionally the code used to link the biographies to Dbpedia is also available in the same git repository, as is the code used to scrape and extract information from each biography.

[12] `https://github.com/munnellg/ODNB_DIB_Dataset`

The ontologies differ from work such as that of Christopher Yocum in that they have been specifically developed to address the requirements of an EL KB as outlined earlier in this chapter. The ontologies also cover a much wider time span from the first century up to the modern age.

The next chapter will combine the work in this chapter with the findings of Chapter 3 in order to develop an effective EL method for CH collections.

Chapter 5

Developing an Method for Linking With Multiple Knowledge Bases

"It's not always easy and sometimes life can be deceiving. I'll tell you one thing, it's always better when we're together."

— Jack Johnson, *Better Together*

5.1 Overview

Given the assessment of available EL systems performed in Chapter 3, the existence of the DIB and ODNB ontologies whose construction was described in Chapter 4, and the awareness of Carmen Brando's proposed approach to perform EL with respect to multiple KBs [39, 9], the purpose of the work presented in this chapter was to investigate the potential for using the previous lines of work to construct an EL system which might be equipped to facilitate better EL performance on specialised CH collections.

The intuition behind the investigated linking method is that a symbiotic relationship can exist between specialised and more general KBs. Information from general KBs will compensate for the sparsity of information in specialised KBs, while the specialised KBs will prevent the EL system from attempting to link mentions to entities that are beyond the scope of the corpus being annotated. In order for this approach to be implemented, the EL system must be capable of recognizing equivalences between entities in its KBs and subsequently resolving these equivalences to some normalised form.

In order to pursue this research, a new EL system designated Cultural Heritage Entity Linker (CHEL) was designed and implemented. This chapter will describe the design details of CHEL, and will report the results of several evaluations which make use of multiple KBs to annotate a variety of content types. For the sake of clarity, CHEL was designed to be an EL system according to this thesis' definition of EL. Hence NER will not be included as a task in the evaluations conducted.

Evaluations are performed on three different content types: the 1641 Depositions corpus, whose construction was described in Chapter 3, a collection of French literary criticisms written by Albert Thibaudet, and a subset of content curated by Europeana[1]. These three collections individually highlight different challenges faced when performing EL in the context of CH.

5.2 Indexing Multiple Knowledge Bases

As was the case with REDEN, CHEL was designed to perform EL with respect to multiple KBs simultaneously. Unlike REDEN, which attempts to resolve the RDF graphs for candidate referents during the referent selection process, CHEL generates a single unified graph of entities during the construction of the KB index.

Mappings which indicate equivalence between two URIs from separate KB sources are passed to CHEL before indexing of the KB begins. Trivially, these equivalences can be derived from properties such as `owl:sameAs` predicates. For entities that are referenced in multiple KBs, CHEL generates a Globally Unique Identifier (GUID) which acts as a unifying identifier across all KBs. This GUID is purely for internal use and is not returned by the system after referent selection. Rather, CHEL is configured to return the URI of the chosen referent from a descending priority list of KBs as specified by the user.

CHEL uses Lucene to create a search index over triples in the source KBs. As each triple is added to the search index, both the subject and the object are checked against the equivalence mappings to see if either or both have been assigned a GUID by CHEL. If such a mapping exists, then the subject or object URI in the triple is replaced with the GUID as appropriate before the triple is added to the search index. In order to facilitate retrieving the original triple from the index, the original URI and all equivalent URIs are also added to the search index. This allows CHEL to filter which sources are used during the EL process and which may be returned as referents.

An example of an entry in the CHEL search index is given in Figure 15. The triple in question indicates that a relationship exists between a woman named Máire O'Brien and a man named Henry Ireton. The original triple from the ontology would have been of the form:

```
<dib:Maire_O'Brien_a6486> <dbo:related> <dib:HenryIreton_a4216> .
```

Máire only exists in the `http://adaptcentre.ie/dib/` KB source (hereafter known as the DIB ontology), which means that she has not received a GUID. The values of her `subjectURI`, `subjectId` and `subjectEquivalents` fields are all equal to her original URI in the DIB ontology.

Henry Ireton exists in both the DBpedia ontology and the DIB ontology. His `objectId` is set to a GUID which identifies all instances of Henry Ireton across all KBs. His `subjectURI` is set to the URI present in the triple, indicating which instance of Henry Ireton in the KB formed the original triple. Finally the

[1] `https://app.assembla.com/spaces/europeana-r-d/documents?folder=58725383`

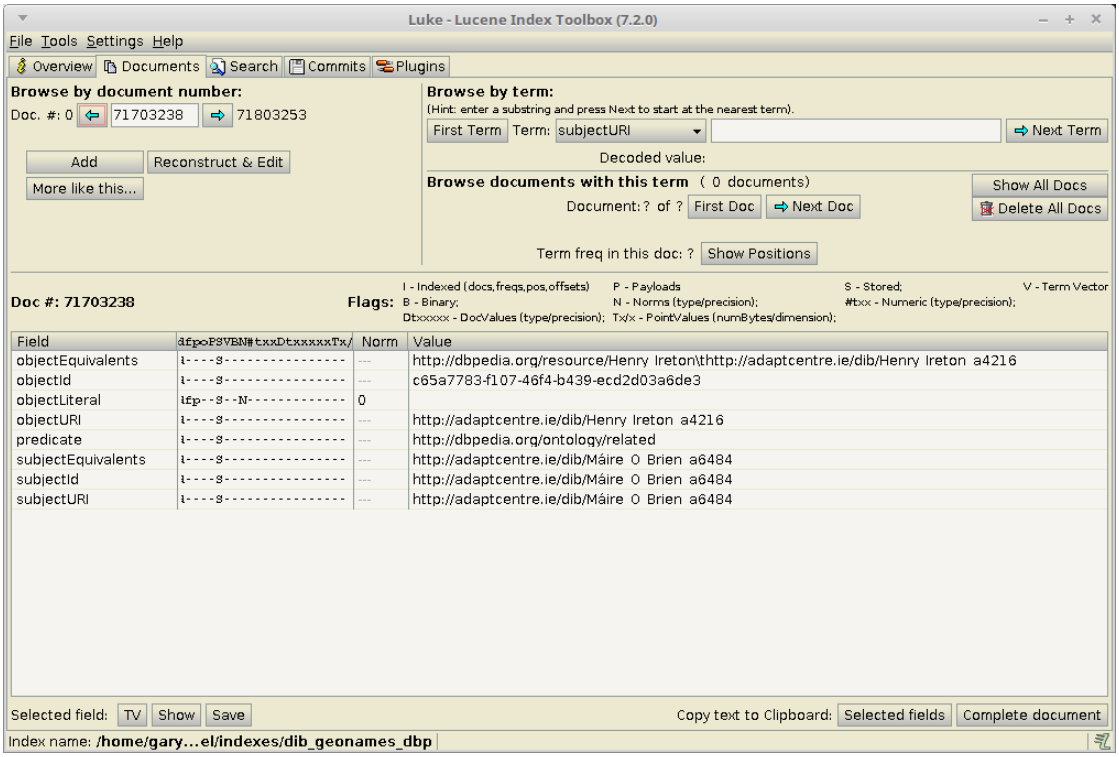


Figure 15: Analysis of CHEL KB search index with Luke

`subjectEquivalents` field is used to store all URIs that may be used to refer to Henry Ireton. While this could be obtained from the GUID mappings datastructure, storing them in the search index reduces the number of queries required to retrieve information during the linking process.

Note that the `objectLiteral` field is used to store surface forms which are queried during candidate selection.

The advantage of constructing the KB in this manner is that it allows the EL system to merge entities that are present in several KB sources, which can help to identify direct relationships between entities that can be difficult to identify when using typical EL indexing approaches.

Consider, as an illustrated example, the diagram presented in Figure 16. The diagram shows entites two KB sources. One source is coloured red, and the other blue. The entities “Charles”, “Offaly” and “Florence” are present in all three KB sources. However, in the blue KB source, only Charles has a direct connection to Offaly. In the red KB source, only Florence has a direct connection. The underlining of Charles and Florence in the red KB source indicates that these two entities are being considered as candidate referents for a mention. Offaly is not being considered as a referent, but it is clearly an important, direct, link between Charles and Florence.

State of the art EL indexing approaches will index these as six separate entities. When analysing the relationships between entities, this implies that there are four degrees of separation between the Florence$_{red}$ entitiy and the Charles$_{red}$ entity i.e. (Florence$_{red}$ $\rightarrow$ Offaly$_{red}$ $\rightarrow$ Offaly$_{blue}$ $\rightarrow$ Charles$_{blue}$ $\rightarrow$ Charles$_{red}$).

In the case of REDEN, the entities are still indexed as six separate entities. However, at runtime, REDEN will merge the Charles and Florence entities across the red and blue KB sources. Offaly will still be split between two entities. This implies that there are three degrees of separation between Florence and Charles (Florence $\rightarrow$ Offaly$_{red}$ $\rightarrow$ Offaly$_{blue}$ $\rightarrow$ Charles), when in reality there are only two degrees of separation.

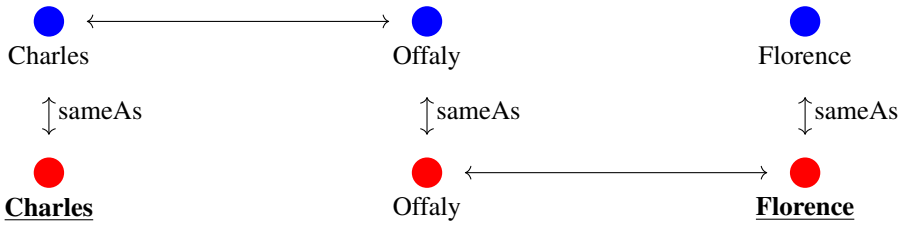


Figure 16: Illustrated example resolving information across multiple KBs

CHEL's novel approach to constructing a KB for EL enables an EL system to correctly resolve the entities in this graph, recognising that there should only be three entities, not six, and that there are only two degrees of separation between Charles$_{red}$ and Florence$_{red}$. Furthermore, this can be achieved with very little modification to the actual referent selection process. The GUIDs simply act as IDs for candidates which are weighted as per normal depending on the EL method used. After the weighting of the graph is complete, a referent from any of the chosen target KBs can simply be extracted by examining the content of the `subjectEquivalents` field associated with the selected referent.

Unlike REDEN which is restricted to the use of a single reference KB, this method facilitates the use of multiple different reference KBs with the `subjectURI` and `objectURI` fields providing provenance information with regards to the source of the entity. If the source is not one of the reference KBs, then the entity is not a valid candidate during candidate selection.

Alternatively, the condition could be imposed that either the `subjectEquivalents` or `objectEquivalents` field must contain at least one URI from the reference KB. This would allow an entity to be found through information outside the reference KB, but only considered as a candidate if it is present in some way in the reference KB.

5.3 Candidate Selection

Entity mentions are executed in turn as queries against the search index whose construction was described in Section 5.2. CHEL is configured by the user to use a subset of available KBs as reference KBs. This means that candidates will only be considered if they are represented in one of chosen reference KBs.

For each valid candidate extracted from the index, CHEL computes the string similarity between the surface form of the candidate and the mention that is the subject of the query. Similarity is computed using Monge-Elkan, whose implementation was discussed in Chapter 4. Once the similarity between all candidates and the original mention have been computed in this manner, the candidates are ranked according to the similarity between the surface form and the mention.

Only the candidates with the most similar surface forms are chosen as candidate referents at this stage in the EL process. In other words, the maximum Monge-Elkan similarity is extracted from the list of candidates and any candidate whose similarity falls below this value is removed from consideration as a referent.

A hard threshold is also applied. If the maximum similarity obtained from the set of candidates is below this threshold, then all candidates retrieved from the index are rejected. This filtered list of candidate referents is returned by the candidate selection process to be considered as referents for the mention.

5.4 Referent Selection

The evaluation in Chapter 3 demonstrated AGDISTIS' efficacy when annotating the depositions, but it was shown that this was due to the candidate selection process rather than due to HITS' scoring. Even so, the choice was made to continue using HITS as it was difficult to affirm or disregard its utility based on the findings of that experiment. Given AGDISTIS' overall performance in the task it was preferable to retain its referent selection method until such time as it could be more adequately assessed.

However, the novel indexing method now enables HITS to be applied to a graph that is constructed across multiple information sources. The internal GUIDs are used to uniquely identify vertices which may be composites of references to an entity obtained from several different sources. After the weighting of the graph is complete, a referent from any of the chosen target KBs can simply be extracted by examining the content of the vertex.

The candidates obtained during the candidate selection process are added to a weighted, directed graph and Breadth First Search (BFS) is run in order to find relationships between entities in the graph. HITS is executed to score each of the candidates and identify which of the available entities may be chosen as a referent for a mention.

HITS is a link-based weighting algorithm used to score vertices in a graph according to two different classes of vertex, namely hubs and authorities. The definition of hubs and authorities is circular. A good hub points to many good authorities, and a good authority is pointed to by many good hubs. For a given vertex in a graph v, this circular definition of hub score $h(v)$ and authority score $a(v)$ yields the following formulation:

$$h(v) = \sum_{v \mapsto \bar{v}} a(\bar{v})$$

$$a(v) = \sum_{\bar{v} \mapsto v} h(\bar{v})$$

The hub score of a vertex is equal to the sum of the authority scores of all vertices that it points to, and the authority score of a vertex is equal to the sum of the hub scores of all vertices that point to the vertex. On initialisation, the algorithm sets the hub and authority scores of all vertices to 1. HITS then proceeds iteratively, computing updated hub and authority scores for each vertex in the graph.

Typically HITS is mentioned in the context of IR for online web content. After a set of documents is returned based on a user's query, HITS is used to score the set of retrieved documents. The hub score can be used to find pages that, perhaps, represent indexes of websites that are relevant to the query while the authority score helps to identify individual pages that may satisfy the user's information need.

In the context of EL, HITS may be used as a substitute for the coherence function $\psi(\Gamma)$ in Ratinov's formal definition of an EL system (discussed previously in Chapter 2):

$$\Gamma^* = \underset{\Gamma}{\operatorname{argmax}} \sum_{i=1}^{|M|} \left[\phi(m_i, \gamma_i) + \psi(\Gamma)\right]$$

The intuition is that good, coherent choices for candidates will manifest themselves as good hubs and/or authorities within a large sub-graph of the overall graph of candidates, while poorer candidates will form smaller, isolated clusters which have low hub and authority scores. A score for the quality of a candidate in the graph may be obtained by averaging its hub and authority scores. After running HITS on the graph, the candidate with the highest score for each mention is selected as the chosen referent.

In addition to specifying a reference KB which determines where candidates entities may be drawn from, the user may also specify a target KB which indicates the KB that should be used to identify the final chosen referents. If no output KB is specified, then CHEL will simply return the URI of the chosen referent. However, if the chosen referent does not exist in the target KB, then CHEL will return a NIL annotation.

5.5 Including REDEN in Evaluations

To the knowledge of this researcher, REDEN is the only other EL system which implements a multi-KB approach to EL. It was therefore deemed important that it be included in evaluations.

In their original work, Brando et al. used a series of metrics inspired by those proposed by Hachey to evaluate REDEN. If given a suitably formatted input file, REDEN will perform EL on the content and will

then evaluate itself, producing a file documenting its results as output. For the purposes of consistency in this thesis, the decision was made to attempt any comparative evaluations involving REDEN using GERBIL.

While a beta version of REDEN's online interface is available[2], at present it will only accept inputs in TEI format [55]. This is to be expected as REDEN was specifically designed to process documents that have been annotated according to the TEI standard. However, as a consequence REDEN cannot be evaluated using GERBIL, as GERBIL uses NIF [49] as the format for communicating the contents of a corpus with EL systems.

In order to circumnavigate this problem, an intermediate service was written which sits between GERBIL and a deployment of REDEN. Note that the REDEN deployment is not REDEN Online, but rather a version of REDEN's desktop app. The intermediate service accepts inputs in NIF format and converts the documents to their equivalent in TEI, which is then written to disk. In truth the "equivalent TEI" is little more than an XML document which adheres to structural properties that REDEN assumes are present in the input. Hence this is not a true conversion, but it is sufficient for purpose.

The intermediate service calls REDEN which processes the documents and performs EL. After REDEN terminates, the intermediate service retrieves REDEN's TEI output, converts the documents back to NIF and sends the result back to GERBIL.

The conversion from NIF to TEI and vice-versa is a simple one-to-one mapping between elements. A single context[3] in NIF format corresponds to a single TEI `div` element. The string content of the context is converted to a `div` and entity mentions are marked using `entity` tags. A `uri` attribute is added to each `entity` element which stores the URI of the phrase in NIF. This makes it easier to map the TEI document back to the original NIF document. Figure 17 provides a sample NIF context with its corresponding XML/TEI equivalent.

The TEI generated is clearly little more than a skeleton XML document, but this is sufficient for REDEN to identify the entity mentions in the input request and to perform the EL task. Similarly the conversion back to NIF is a direct one-to-one mapping which searches for `ref_auto` attributes (the attribute in which REDEN stores its output) in the XML document.

REDEN outputs URIs for referents from all KBs in which the chosen referent is represented. However, GERBIL only expects a single annotation per mention in the result document. The intermediate service is configured to select a referent URI from those returned by REDEN using a descending priority list of KB namespaces. If none of the URIs returned by REDEN are in the list of ranked namespaces, then the intermediate service returns the first annotation supplied by REDEN. If REDEN does not apply an annotation to a mention, then the intermediate service returns NIL.

[2] `http://obvil-dev.paris-sorbonne.fr/reden/RedenOnline/site/input-tei.html`
[3] `http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Context`

```
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-↩
    core#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.com/document#char=16,26> a nif:Phrase,
        nif:RFC5147String,
        nif:String ;
    nif:anchorOf "La Bruyere"^^xsd:string ;
    nif:beginIndex "16"^^xsd:nonNegativeInteger ;
    nif:endIndex "26"^^xsd:nonNegativeInteger ;
    nif:referenceContext <http://example.com/document#char=0,65> .

<http://example.com/document#char=0,65> a nif:Context,
        nif:RFC5147String,
        nif:String ;
    nif:beginIndex "0"^^xsd:nonNegativeInteger ;
    nif:endIndex "65"^^xsd:nonNegativeInteger ;
    nif:isString """Quoi qu'en dise La Bruyere, il peut en naitre des chefs↩
        -d'oeuvre."""^^xsd:string .
```

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<TEI>
    <text>
        <body>
            <div>Quoi qu'en dise <entity uri="http://example.com/document#↩
                char=16,26">La Bruyere</entity>, il peut en naitre des ↩
                chefs-d'oeuvre.</div>
        </body>
    </text>
</TEI>
```

Figure 17: Sample mapping from NIF document (top) to TEI (bottom) and vice-versa

In the process of testing REDEN it was also found that some of the APIs it uses to harvest data from external sources e.g. Bibliotheque Nationale De France (BnF), have changed since REDEN was originally written. Most notably, when retrieving the RDF document for a BnF entity, the URI of the entity no longer downloads the corresponding RDF/XML. Rather, each entity in BnF has an associated `foaf:Page` attribute which can be used to retrieve its corresponding RDF. In order to facilitate this change, REDEN's approach to downloading online resources was modified to route all requests through the intermediate service which communicates with GERBIL. The intermediate service identified the appropriate link to satisfy the request and routed the result of querying the API back to REDEN.

Due to the implementation of this service, it was possible to perform tests on REDEN using GERBIL.

Hachey's measures have proved to be useful for performing a deeper analysis of EL systems. Hence REDEN was also configured to save the outputs of its candidate selection process to a file which could be examined to measure the efficacy of the candidate selection process.

REDEN does not provide specific version numbers, but for the purposes of the evaluations performed in this chapter, note that the repo was cloned in Summer 2017.

5.6 Deploying GERBIL

The experiments in this chapter use a number of KBs which are not typically indexed by the online deployment of GERBIL. In order to carry out the evaluations using multiple KBs, a local deployment of GERBIL was set up and configured.

GERBIL uses a series of Lucene search indexes in order to determine which entities are considered to be InKB and which constitute Emerging Entities. A `sameAs` index is used to establish equivalence between URIs. This allows GERBIL to evaluate, for example, a system which uses YAGO as a KB on a corpus annotated with DBpedia URIs. When downloaded and run for the first time, GERBIL will automatically download the necessary triples and set up these common indexes.

For the more specialised KBs used in the following experiments, a new `sameAs` index was created based on known equivalences between the various KB sources. These equivalences were based on `skos:exactMatch` and `owl:sameAs` predicates present in the KB sources.

Additionally the entities in the KBs were added to Lucene indexes which GERBIL was configured to access during evaluations. GERBIL is provided with tools which perform the construction of this entity index given an input series of triples in Notation3 format. Aside from the addition of these new indexes, GERBIL was run using all default settings from the deployment available on Github[4].

[4] `https://github.com/dice-group/gerbil`

5.7 Baseline Depositions Evaluation

An initial baseline assessment of CHEL was performed using the 1641 depositions as a test corpus and only DBpedia as the KB. The purpose of the investigation was to establish a measure of how well CHEL would perform under the experimental conditions described in Chapter 3. As before, GERBIL was used as a testing platform and was configured to perform a D2KB evaluation. The local deployment of GERBIL described in Section 5.6 was used to perform the test.

In addition to performing an evaluation with GERBIL, CHEL was also configured to report its candidate referent sets after the candidate selection process. This facilitates an additional level of scrutiny using Hachey's evaluation measures. It is interesting and highly useful to consider the performance of an EL system at this intermediate stage, rather than treating the entire process as a black box and simply evaluating based on the final results. The use of Hachey's measures can either shed light on reasons behind the scores produced by GERBIL, or even potentially create a conflict between scores. If there is disagreement between the two evaluation approaches, it would suggest some level of experimental error.

Experimenting with different configurations of CHEL in this test showed that it performed best in the Standard GERBIL evaluation when the threshold on Monge-Elkan was set to 0.98. This makes the system rather conservative in its selection of candidate referents during the candidate selection phase. The results of the experiment are reported in Tables 5.1 and 5.2. For ease of comparison, the results from the earlier AGDISTIS evaluation are presented in tables 5.3 and 5.4.

Evaluation Type	**Macro F1**	**Macro Precision**	**Macro Recall**	**Micro F1**	**Micro Precision**	**Micro Recall**
Standard	**0.6083**	**0.6083**	**0.6083**	**0.6269**	**0.6269**	**0.6269**
InKB	0.3064	0.3642	0.2782	0.3293	0.3901	0.2850
EE	**0.7542**	**0.7073**	**0.8340**	**0.7572**	**0.6991**	**0.8258**
GSInKB	0.2782	0.2782	0.2782	0.2850	0.2850	0.2850

Table 5.1: Results GERBIL evaluation on CHEL using DBpedia as KB.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.4669	1.6712	**0.3623**	**0.4811**	**0.3165**	**0.7542**	**0.7073**	**0.8340**
Macro	0.4729	1.6099	**0.3713**	**0.4397**	**0.3212**	**0.7572**	**0.6991**	**0.8258**

Table 5.2: Results for Hachey's measures on CHEL using DBpedia as KB.

The results of the evaluation are largely unsurprising. They are more or less exactly what is expected given what was learned from the investigation of AGDISTIS in Chapter 3. CHEL has achieved marginally higher scores in the Standard and EE evaluations and a marginally lower score in the InKB and GSInKB

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.5979	0.5979	0.5979	0.6052	0.6052	0.6052
InKB	**0.3395**	**0.4589**	**0.3063**	**0.3557**	**0.4040**	**0.3177**
EE	0.7189	0.6858	0.7840	0.7326	0.6869	0.7847
GSInKB	**0.3063**	**0.3063**	**0.3063**	**0.3177**	**0.3177**	**0.3177**

Table 5.3: Results GERBIL evaluation on AGDISTIS using DBpedia as KB.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	1.5292	5.3577	0.3161	0.3796	0.2708	0.7068	0.6501	0.7743
Macro	1.5550	8.1619	0.3324	0.4147	0.2773	0.7148	0.6640	0.7742

Table 5.4: Results for Hachey's measures on AGDISTIS using DBpedia as KB.

evaluations. The difference can be ascribed to the fact that CHEL was configured during the experiment to have a high threshold for similarity between a surface form and a mention before an entity in the KB might be considered as a referent. CHEL was set up to be more strict than AGDISTIS leading to a higher overall score in the standard evaluation given the prevalence of NIL annotations in the 1641 depositions corpus.

As was the case with the AGDISTIS evaluation it can be seen from Hachey's metrics that the scores achieved in GERBIL can largely be ascribed to the candidate selection process, rather than a discerning referent selection process. On average the candidate set for a given mention (assuming at least one candidate was found) contains only one or two entities, as indicated by a $\mu_{C \text{ not NIL}}$ score of approximately 1.6 in both the micro and macro evaluations. This means that HITS tends to have either a very simple binary choice to make between two referents, or else it simply chooses the only referent it is given for a particular mention.

Once again, the P_{NIL}, R_{NIL} and $F1_{NIL}$ scores are considerably higher than their counterparts for the candidate set, showing that the referent selection process is reasonably capable of identifying when a mention should not have a referent. However, the low scores for P_C, R_C and $F1_C$ show that it often fails to recognise the existence of a referent in the KB given that one has been identified in the gold standard. Respectively, the $F1_C$ score in the micro and macro measures are 0.3623 and 0.3713. Less that 40% of entities in the gold standard that should have a referent were even identified in the KB before HITS began to analyse the graph.

Note that the scores for the EE task from GERBIL and the values of P_{NIL}, R_{NIL} and $F1_{NIL}$ in Hachey's metrics are equal for CHEL. This is to be expected under these experimental conditions as the candidate

selection process is effectively what determines whether or not a NIL annotation should be applied to a mention. In the case of AGDISTIS the values differ slightly as this experiment needed to be run on two different deployments of AGDISTIS, one being the "official" AGDISTIS deployment provided by AKSW and the other being a local deployment that was modified to generate a file that could be used to compute Hachey's scores. This results in a slight difference between the two sets of results.

With this baseline evaluation completed, further experimentation examined the effects of introducing other, more specialised KB sources into the EL process.

5.8 Multiple Knowledge Base Evaluation

An evaluation was carried out to determine whether or not CHEL's performance on the depositions would be improved by the use of multiple KBs during the linking process. Annotations currently applied to the depositions mean that mentions belong to one of two classes of entities, namely people and places.

The DIB and ODNB KBs constructed in Chapter 4 help with the problem of identifying people in the depositions. However, a separate source is required to deal with the problem of identifying locations. Two notable sources identified in Chapter 2 are Geohive and GeoNames. Due to the fact that the evaluation corpus was annotated using DBpedia URIs and the need for references between the various sources which make up the KB, the decision was made to use GeoNames rather than Geohive to identify locations. While Geohive has better coverage of Irish locations, it does not have any links to DBpedia which means the two sources cannot be integrated with one another. A subset of GeoNames pertaining to Ireland was downloaded and used in the KB. Equivalence between entities in DBpedia and GeoNames was established using the `geonames_links_en.ttl` dataset from DBpedia.

The use of ODNB and DIB were considered separately as these two sources do not directly link to each other at the present time. Both KBs were filtered to contain entities relevant to the time period around the rebellion using a hard threshold which dictated that an entity must have been born in the latter half of the 16^{th} century or the first half of the 17^{th} century.

CHEL was configured to use DIB/ODNB and GeoNames as the reference KBs but to return the final chosen referent using a URI from DBpedia due to the fact that the evaluation corpus was annotated with DBpedia URIs.

As usual, GERBIL was used to conduct the evaluation as a D2KB experiment. The results of the evaluation using DIB, GeoNames and DBpedia are presented in Tables 5.5 and 5.6, while the results of the experiment using ODNB, GeoNames and DBpedia are presented in Tables 5.7 and 5.8.

Using multiple KBs in the manner described yields an appreciable improvement in the quality of results returned by CHEL. It can be seen that the performance of the service across all categories of the GERBIL evaluation have increased, with the scores in the standard evaluation jumping from $F1_{macro} = 0.6083$

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	**0.7833**	**0.7833**	**0.7833**	**0.8013**	**0.8013**	**0.8013**
InKB	**0.6344**	0.8631	**0.5242**	**0.6275**	0.8496	**0.4974**
EE	**0.8588**	**0.7806**	**0.9775**	**0.8563**	**0.7629**	**0.9756**
GSInKB	**0.5242**	**0.5242**	**0.5242**	**0.4974**	**0.4974**	**0.4974**

Table 5.5: Results of GERBIL evaluation on CHEL using DBpedia, GeoNames and DIB as KB.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.7312	2.4207	**0.5621**	**0.6552**	**0.4922**	**0.8199**	**0.7612**	**0.8885**
Macro	0.7200	2.2844	**0.5779**	0.6899	**0.5205**	**0.8254**	**0.7762**	**0.9032**

Table 5.6: Results for Hachey's measures on CHEL using DBpedia, GeoNames and DIB as KB.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.7729	0.7729	0.7729	0.7933	0.7933	0.7933
InKB	0.6253	**0.8666**	0.5109	0.6221	**0.8774**	0.4819
EE	0.8462	0.7623	0.9732	0.8411	0.7433	0.9686
GSInKB	0.5109	0.5109	0.5109	0.4819	0.4819	0.4819

Table 5.7: Results of GERBIL evaluation on CHEL using DBpedia, GeoNames and ODNB as KB.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.7333	2.5324	0.5542	0.6619	0.4767	0.8025	0.7390	0.8780
Macro	0.7215	2.5228	0.5725	**0.6997**	0.5072	0.8111	0.7561	0.8986

Table 5.8: Results for Hachey's measures on CHEL using DBpedia, GeoNames and ODNB as KB.

in the previous experiment to $F1_{macro} = 0.7833$ for DIB and $F1_{macro} = 0.7729$ for ODNB. The increase in the GSInKB task indicates that, while the accuracy of linking entities is still lower than may be considered desirable, it is considerably higher than was the achieved when using DBpedia in isolation.

Examining Hachey's measures shows that the use of multiple KBs has also had a noticeable effect on the candidate selection process. Both the scores for $F1_C$ and $F1_{NIL}$ have increased significantly, suggesting that the inclusion of the new KBs has improved both CHEL's ability to identify that the correct referent exists in the KB and its ability to abstain from annotating where appropriate.

Overall, the use of multiple KBs appears to have had a dramatic and positive impact on the performance of this graph-based EL system when applied to the 1641 depositions.

5.8.1 Comparison With REDEN

For comparative purposes, it was intended to carry out an evaluation using REDEN under the same conditions outlined in the previous section. However, REDEN can only use one reference KB per problem. As the depositions experiment requires the use of both GeoNames and either DIB or ODNB to annotate all classes of entity, this limitation meant that REDEN could not be fairly assessed under the previous experimental setup.

The decision was made to split the comparison into two separate evaluations based on entity type. An evaluation would be performed based on locations in the depositions using GeoNames and DBpedia, and a second evaluation would be performed with respect to people using ODNB and DIB with DBpedia. REDEN communicated with GERBIL during this evaluation via the service described in Section 5.6.

The results of these individual experiments are presented across the tables on the following pages. Tables 5.9, 5.10, 5.11 and 5.12 provide the results for the experiment with GeoNames and DBpedia. Tables 5.13, 5.14, 5.15 and 5.16 provide the results for DIB and DBpedia. Tables 5.17, 5.18, 5.19 and 5.20 provide the results for ODNB and DBpedia.

Due to the large volume of data presented in the preceeding tables, a very brief summary of the results is provided in Table 5.21. This table only presents the Macro F1 scores of the Standard evaluation for each EL system on the 1641 depositions for different entity types and different combinations of KB source. Note that this summary table also omits Hachey's measures. Hence, while it provides a very high level view of how each system performed in respective evaluations, it does not give much insight into the source of an EL system's success or failure. For that information, it is recommended that the reader view the full results tables.

An N/A entry in Table 5.21 indicates that the EL system was not used for a particular evaluation because it was not possible for the system to carry out the task. An — symbol indicates where data was not gathered.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	**0.7118**	**0.7118**	**0.7118**	**0.7383**	**0.7383**	**0.7383**
InKB	**0.7368**	**0.9025**	**0.6463**	**0.7308**	0.9135	**0.609**
EE	**0.7171**	**0.627**	0.9498	**0.6869**	**0.544**	0.9315
GSInKB	**0.6463**	**0.6463**	**0.6463**	**0.609**	**0.609**	**0.609**

Table 5.9: Results of GERBIL evaluation on CHEL using DBpedia and GeoNames as KB and locations from 1641 depositions.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	1.5153	2.5328	**0.6416**	0.6861	**0.6026**	**0.5333**	**0.4783**	0.6027
Macro	1.3999	2.5436	**0.6684**	**0.7170**	**0.6418**	**0.5789**	**0.5659**	0.6658

Table 5.10: Results for Hachey's measures on CHEL using DBpedia and GeoNames as KB and locations from 1641 depositions.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.5415	0.5415	0.5415	0.5537	0.5537	0.5537
InKB	0.4042	0.7274	0.303	0.4952	**0.963**	0.3333
EE	0.5742	0.4381	**0.9952**	0.5806	0.4114	**0.9863**
GSInKB	0.303	0.303	0.303	0.3333	0.3333	0.3333

Table 5.11: Results of GERBIL evaluation on REDEN using DBpedia and GeoNames as KB and locations from 1641 depositions.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.5153	1.8438	0.4727	**0.8125**	0.3333	0.5714	0.4121	**0.9315**
Macro	0.4533	1.5615	0.3955	0.6623	0.3030	0.5636	0.4360	**0.9431**

Table 5.12: Results for Hachey's measures on REDEN using DBpedia and GeoNames as KB and locations from 1641 depositions.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.8452	0.8452	0.8452	0.8234	0.8234	0.8234
InKB	0.3105	0.3646	0.3031	0.1176	0.2143	0.0811
EE	0.8988	0.86	0.9698	0.9272	0.8824	0.9767
GSInKB	0.3656	0.3656	0.3656	0.0811	0.0811	0.0811

Table 5.13: Results of GERBIL evaluation on CHEL using DBpedia and DIB as KB and people from 1641 depositions.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.1032	1.2381	**0.2414**	**0.3333**	**0.1892**	0.9103	0.8788	0.9442
Macro	0.0843	0.5391	**0.1103**	**0.1719**	**0.1062**	0.8842	0.8583	0.9447

Table 5.14: Results for Hachey's measures on CHEL using DBpedia and DIB as KB and people from 1641 depositions.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	**0.8651**	**0.8651**	**0.8651**	**0.8419**	**0.8419**	**0.8419**
InKB	**0.3343**	**0.3854**	**0.3292**	**0.1702**	**0.3636**	**0.1111**
EE	**0.9095**	**0.8661**	**0.9875**	**0.9365**	**0.888**	**0.9907**
GSInKB	**0.3917**	**0.3917**	**0.3917**	**0.1111**	**0.1111**	**0.1111**

Table 5.15: Results of GERBIL evaluation on REDEN using DBpedia and DIB as KB and people from 1641 depositions.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.0516	1.0000	0.1600	0.3077	0.1081	**0.9295**	**0.8828**	**0.9814**
Macro	0.0455	0.3750	0.0783	0.1354	0.0688	**0.9014**	**0.8626**	**0.9760**

Table 5.16: Results for Hachey's measures on REDEN using DBpedia and DIB as KB and people from 1641 depositions.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.8333	0.8333	0.8333	0.8216	0.8216	0.8216
InKB	0.1364	0.1875	0.1313	0.0455	0.1429	0.027
EE	0.8913	0.8435	0.9754	0.9087	0.8531	0.9721
GSInKB	0.3187	0.3187	0.3187	0.027	0.027	0.027

Table 5.17: Results of GERBIL evaluation on CHEL using DBpedia and ODNB as KB and people from 1641 depositions.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.0952	1.7143	0.0392	0.0714	0.0270	0.8918	0.8487	0.9395
Macro	0.0676	0.5625	0.0104	0.0312	0.0063	0.8794	0.8402	0.9531

Table 5.18: Results for Hachey's measures on CHEL using DBpedia and ODNB as KB and people from 1641 depositions.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.8333	0.8333	0.8333	0.8216	0.8216	0.8216
InKB	0.1364	0.1875	0.1313	0.0455	0.1429	0.027
EE	0.8913	0.8435	0.9754	0.9087	0.8531	0.9721
GSInKB	0.3187	0.3187	0.3187	0.027	0.027	0.027

Table 5.19: Results of GERBIL evaluation on REDEN using DBpedia and ODNB as KB and people from 1641 depositions.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	0.0952	1.7143	0.0392	0.0714	0.0270	0.8918	0.8487	0.9395
Macro	0.0676	0.5625	0.0104	0.0312	0.0063	0.8794	0.8402	0.9531

Table 5.20: Results for Hachey's measures on REDEN using DBpedia and ODNB as KB and people from 1641 depositions.

	CHEL	REDEN	AGDISTIS
KB Sources	1641 Depositions People & Places		
DBpedia	**0.6083**	—	0.5979
DBpedia, GeoNames, DIB	0.7833	N/A	N/A
DBpedia, GeoNames, ODNB	0.7729	N/A	N/A
	1641 Depositions People		
DBpedia, DIB	0.8452	**0.8651**	N/A
DBpedia, ODNB	0.8333	0.8333	N/A
	1641 Depositions Places		
DBpedia, GeoNames	**0.7118**	0.5415	N/A

Table 5.21: Summary of results shown in Tables 5.1 to 5.20

Comparing REDEN and CHEL based on locations and people separately yields some interesting insights into the results of the evaluation performed in Section 5.8.

On the challenge of linking locations, CHEL performs significantly better than REDEN. It can be seen from the results in Tables 5.11 and 5.12 that REDEN's lower scores can be ascribed to the fact that it struggles to identify the correct referent during the candidate selection process. In Table 5.12, REDEN has a very high R_{NIL} score, but a low P_{NIL} score, showing that it liberally and often inappropriately applies NIL annotations to surface forms in the depositions.

CHEL, however strikes a good balance between annotating locations and abstaining from annotating. This is found to be due to its more lenient string similarity threshold when comparing the spelling of mentions to the surface forms of candidates in the KB. REDEN requires exact string matches which are too strict for the noisy language found in the depositions.

The performance of REDEN and CHEL when using ODNB and DIB are disappointing. While the $F1$ scores in the standard evaluation for both the macro and micro measures are extremely high, the GSInKB and InKB scores are quite low. Recalling that the InKB evaluation measures how well the EL system performs when applying URIs to all mentions, and the GSInKB evaluation examines how well the EL system performs when applying annotations to mentions that should have URIs, this suggests that neither EL system performs well at finding correct referents in the specialised KB. There are a variety of conditions which explain this.

The difference in performance between ODNB and DIB is due in part to the focus of the respective

collections of biographies. ODNB is, of course, focused on significant figures in the history of the UK while DIB is focused on Irish people. This means that ODNB has slightly poorer coverage than DIB with respect to the Irish rebels. Most significantly, the English king, Charles I, is mentioned in ODNB but not DIB.

Another difficulty can be ascribed to entities that are missing from both the ODNB and DIB ontologies. Notably figures such as the King of Spain (Philip IV) and the Catholic Pope (Pope Urban VIII) will not be found in biographies that describe significant English/Irish people. One possible solution to this problem would be to directly query DBpedia as a KB if no referent is found in the specialised KB. Whether or not this is an appropriate thing to do depends on precisely what is the intended role of the reference KB.

If it is intended that the reference KB strictly limits the general KB to a subset of entities, then circumnavigating the reference KB is something that should not be done. However, a softer view of the reference KB could be taken where it represents the domain knowledge that is highly specific, but known to be incomplete with respect to certain entities. If the domain specific knowledge fails, then falling back on a general source might help to fill in any gaps in the general knowledge of the reference KB.

It can also be seen that some entities are not annotated due to a lack of surface forms in the KB. While an effort was made to generate permutations of surface forms in the ODNB and DIB ontologies, entities such as Charles I are referenced by title rather than by name in the depositions. So, while "Charles I" is a valid surface form in the KB, the surface form "The King" is not associated with this entity. While it is perfectly possible to simply add this surface form which would cause an increase in the performance of the ontologies, this would be cheating in the greater context of the research that was being conducted. Part of the investigation was to use historical data that was automatically harvested from sources used by historical scholars. Hence in their raw form, and even with a sensible sequence of permutations applied to entity names, the coverage of surface forms is not necessarily sufficient.

Hachey's metrics in tables 5.14, 5.16, 5.18 and 5.20 show that the problems identified thus far manifest themselves at the level of the candidate selection process, often preventing the correct referent from being found in the KB.

While the systems did struggle to identify correct referents, it can be seen that the use of DIB and ODNB helped both CHEL and REDEN to attain extremely high scores in the EE task, and subsequently a strong performance in the standard evaluation. This demonstrates how the use of multiple KBs in the manner implemented in both systems helps to control the scope of the EL system. Some of the names in the depositions are extremely common e.g. "Peter" and can easily be linked to an incorrect referent by an overly optimistic EL system. With the use of multiple KBs, this has been held in check.

The performance of REDEN and CHEL is reasonably on par for both DIB and ODNB, with REDEN actually performing marginally better than CHEL on DIB. Both systems achieve exactly the same performance in the evaluation with ODNB. Hachey's measures show that this is due to both systems struggling

to identify the correct referent during the candidate selection process. This is likely due to the issues highlighted previously.

Overall it can be seen that the use of multiple KBs in this context has helped to keep both CHEL and REDEN in check with respect to applying URIs to mentions. There are clear means by which the performance in the InKB and GSInKB tasks may be improved through the expansion of the ODNB and DIB ontologies, or even through the inclusion of another KB source that can help to compensate where these sources are lacking.

5.9 Evaluation on French Literary Criticism

REDEN was originally evaluated using a French literary criticism written by Albert Thibaudet in 1936. Thibaudet's "Réflexions sur la Littérature" are a different class of problem to the depositions in a number of ways. First, the text was written in the 20^{th} century, resulting in mentions of more contemporary entities, and therefore fewer NIL annotations. The text is considerably less noisy than the archaic language found in the depositions with consistent spelling and more natural sentence structure. It is, perhaps, worthwhile mentioning that the Réflexions are written in French. However, this has little impact on the execution of either CHEL or REDEN given that both only consider the spelling of surface forms in order to choose a referent for a mention.

An investigation was carried out on the original Thibaudet corpus in order to see how CHEL would perform compared to REDEN on the task for which REDEN was originally designed.

The corpus contains 3,404 mentions of people, including fictional characters (2,980 if fictional characters are excluded). Of these 3,404 entities, 1916 were found to have a referent which was identified with a URI drawn from either BnF or Identifiants et Référentiels pour l'Enseignement Supérieur et la Recherche (IdRef) in the gold standard. This corpus is distributed with REDEN via the project's GitHub repository and is stored in TEI-XML format. Due to the use of GERBIL for evaluations, the Réflexions were converted to NIF format for the purposes of this experiment.

During the conversion, a `nif:Context` was defined to be a `div` element that was a direct child of the `body` element in the document. This conforms to REDEN's definition of a context as indicated by the default configuration files supplied with the project. Strings of text between tags marked `persName` were marked as mentions in the corresponding context. The result is an evaluation corpus comprised of 28 documents.

The KBs used by both REDEN and CHEL were BnF and the French DBpedia dataset. Entities from BnF were downloaded using a SPARQL query supplied in REDEN's source code. The query downloads instances of entities which are tagged as belonging to the class `foaf:Person` in the bnf ontology, and whose birthday is some time after the year 1900.

It is important to note that part of REDEN's KB construction process involves performing various manipulations on surface forms in the KB. These manipulations are not too dissimilar to those discussed in the construction of the DIB and ODNB KBs in Chapter 4. Forenames are initialed and prefixes such as titles or honorifics are variously inserted and removed generating several permutations of an author's name. Due to REDEN's strict matching on surface forms, this is a necessary step in order to increase the number of surface forms that will be recognised as references to a specific candidate mention. REDEN was allowed to perform these manipulations on BnF before the experiment began. CHEL also used these transformations to increase the set of surface forms that might identify an entity.

Equivalences between BnF and DBpedia were extracted using `skos:exactMatch` and `owl:sameAs` triples found in BnF. This information was also used to construct a `sameAs` index for GERBIL.

The experiment was run as a D2KB experiment using GERBIL.

REDEN and CHEL achieve similar results when tested on Thibaudet. However, REDEN clearly outperforms CHEL on this content type when the values in Tables 5.22 and 5.24 are compared.

The differences in the performance of the candidate selection process in Tables 5.23 and 5.25 is small, but the value for $F1_C$ suggests that CHEL is marginally better at identifying when a referent for a mention exists in the KB, although the difference in performance between CHEL and REDEN in this capacity is almost negligible. Similarly the value for $F1_{NIL}$ suggests that overall CHEL performs better at identifying when a referent does not exist. Seemingly, CHEL's approach to identifying referents in the KB performs better than REDEN in this task.

However, when considering the results of the final EL output for both systems, the GSInKB test performed by GERBIL shows that REDEN is better than CHEL at choosing the correct referent. The extremely similar values for P_C and R_C in Tables 5.23 and 5.25 shows that the candidate selection processes in both systems achieved similar accuracy when identifying the correct referent in the KB. This means that the scores in the GSInKB task are indicative of which graph weighting approach was better suited to selecting the correct referent from the pool of candidates. In this capacity, REDEN's Degree Centrality Measure performs noticeably better than HITS.

Indeed, the scores for μ_C and $\mu_{C\,\text{not NIL}}$ in Tables 5.23 and 5.25 show that, on average, the pool of candidates for each mention was considerably larger than was typically the case when evaluating with respect to the depositions. This shows that the referent selection process had more work to do on this collection and thus had a greater impact on the final results.

Based on the results of this experiment, it can be concluded that Degree Centrality is a better choice of weighting algorithm than HITS for scoring graphs in this type of CH EL problem.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.5358	0.5358	0.5358	0.5122	0.5122	0.5122
InKB	0.5398	0.4391	0.8316	0.5895	0.4734	0.781
EE	**0.3544**	0.8502	**0.2366**	**0.4195**	0.8948	**0.274**
GSInKB	0.8316	0.8316	0.8316	0.781	0.781	0.781

Table 5.22: Results of GERBIL evaluation on CHEL using French DBpedia and BnF as KB and people from Thibaudet.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	6.3231	7.4221	**0.7016**	**0.5634**	0.9295	**0.4195**	0.8948	**0.2740**
Macro	6.3749	7.5571	**0.6190**	**0.5079**	0.9292	**0.3544**	0.8502	**0.2366**

Table 5.23: Results for Hachey's on CHEL using French DBpedia and BnF as KB and people from Thibaudet.

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	**0.5649**	**0.5649**	**0.5649**	**0.5383**	**0.5383**	**0.5383**
InKB	**0.5791**	**0.4689**	**0.8926**	**0.6372**	**0.5083**	**0.8538**
EE	0.3315	**0.9288**	0.2143	0.4025	**0.9357**	0.2564
GSInKB	**0.8926**	**0.8926**	**0.8926**	**0.8538**	**0.8538**	**0.8538**

Table 5.24: Results of GERBIL evaluation on REDEN using French DBpedia and BnF as KB and people from Thibaudet.

	μ_C	μ_C not NIL	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	6.5911	7.5977	0.6958	0.5550	**0.9323**	0.4025	**0.9357**	0.2564
Macro	6.6115	7.6783	0.6133	0.4995	**0.9316**	0.3315	**0.9288**	0.2143

Table 5.25: Results for Hachey's on REDEN using French DBpedia and BnF as KB and people from Thibaudet.

5.10 Evaluation on Europeana Corpus

A third corpus which was identified as potentially interesting for evaluating EL systems in the context of CH was a corpus generated by a Europeana taskforce led by Antoinne Isaac [56]. The task force sought to investigate methods of automatically enriching content indexed by the Europeana portal. The team created an annotated dataset using random samples of data obtained from The European Library (TEL).

The creation of the evaluation corpus involved the use of automatic EL services. A random sample of content from TEL was obtained with the intention of capturing the diverse range of content that is curated there. Approximately 1,000 samples each from 19 different languages were obtained by the task force. These samples were sent out to be automatically annotated by a variety of different EL services. The results of this automatic annotation process were returned to the task force with several different KB sources having been used to identify the referents.

The task force normalised the referents to a common set of vocabularies using GeoNames for locations and DBpedia for other entity types. Given the normalised annotations, the researchers identified agreements between the different EL services, noting where more than one service had supplied a given annotation for a specific mention. These agreements were used to generate annotation sets based on combinations of agreements between different tools. The result was 26 different sets which were randomly sampled to select up to 100 annotations per set. Where a set contained fewer than 100 annotations, the entire contents of the set were sampled. The result was a collection of 1,757 annotations.

The annotations were manually checked for correctness by a team of annotators. Tools were assessed based on the completeness of a match (how much of the appropriate entity mention was tagged) and the accuracy of the supplied URI. Correction of the corpus did not seem to require that the annotators replace an incorrect URI with its correct counterpart. Rather the incorrect entry was simply marked as such, but left with its original annotation. As will be noted later this creates a risk with the Europeana corpus that the entities it contains are somewhat idealised with respect to the EL problem.

It is also worth mentioning that the contexts of the corpus items is often quite short. Usually there is only one mention per context. A graph-based approach like HITS is likely to suffer here, as their most basic assumption that relationships between entities can help to identify a correct subset of referents is invalidated.

The corpus as supplied by the task force is in a spreadsheet with cells in the spreadsheet indicating a content item, a mention within the content and a referent for the mention. These spreadsheets were transformed to NIF format for use with GERBIL. However not all of the 1,757 annotations were suitable for use, given that some had an incorrect URI associated with them.

The task force indicated the severity of an error using a simple scale which marked that an annotator believed an annotation was correct, incorrect or was unsure of the accuracy of the URI. For this evaluation, the dataset was filtered to remove all annotations that had been marked as incorrect, or which the

annotators were unsure about. Of the remaining collection items, it was noted that some had been annotated with Wikidata, UNESKOS and other vocabularies. Collection items which had not been annotated using either DBpedia or GeoNames were removed from the evaluation corpus. The final result was an evaluation corpus containing 1,267 contexts with 1,329 annotations whose referents were indicated by a URI from either DBpedia or GeoNames. This filtering, however, leads to a somewhat idealised corpus, as the entities which are used are entities that EL systems were previously able to correctly identify and annotate. This would suggest that the entities are less challenging for an EL system to annotate than those seen in Thibaudet or the depositions.

The content of the corpus is rather diverse, but is generally more scientific than the content that was handled in the 1641 depositions or Thibaudet. Many of the content items are drawn from fields such as biology and mathematics. This makes the Europeana corpus a very different challenge to either of the previously investigated collections. One of the consequences of the manner in which this corpus was constructed is that it contains no NIL entities. It is expected that an EL service using GeoNames and DBpedia should be able to annotate every single item in the corpus.

An evaluation was performed using DBpedia and GeoNames simultaneously as reference KBs. Unlike the earlier evaluations which used a reference KB to narrow the search of the EL system, this evaluation uses two KBs together in order to cast the search net as wide as possible, allowing both KBs to compensate for each others respective limitations.

The goal was to correctly annotate as many mentions in the corpus using both KBs. GERBIL was configured to recognise the equivalence between a DBpedia URI and a Geonames URI where such an equivalence existed. Hence the goal was not to annotate locations with GeoNames and other content types with DBpedia. Rather, it was simply to examine the effect of attempting to annotate this collection while considering two KBs at the same time.

This evaluation required the simultaneous use of GeoNames and DBpedia as reference KBs which rendered REDEN unusable for the test, as REDEN is limited to the use of a single reference KB. It was therefore not included in this evaluation. Indeed, as of the time of writing we are unaware of any existing EL system other than CHEL that can carry out this task.

As usual, GERBIL was set up to evaluate the results of the annotation process and was configured to perform a D2KB experiment, the results of which are presented in Table 5.26. Hachey's measures are reported in Table 5.27.

Examining the results, what is immediately noteworthy is the performance of CHEL according to Hachey's metrics with regards to the values of P_C, R_C and $F1C$. In both the micro and macro tasks, the scores are extremely high, showing that CHEL's candidate selection process almost always successfully identified the correct referent in the KB and returned it for consideration as a candidate. This means that the lower scores in GERBIL's InKB and GSInKB tasks can be almost entirely ascribed to the use of HITS as a

Evaluation Type	Macro F1	Macro Precision	Macro Recall	Micro F1	Micro Precision	Micro Recall
Standard	0.7607	0.7607	0.7607	0.7618	0.7618	0.7618
InKB	0.7619	0.7622	0.7618	0.7677	0.7747	0.7607
EE	0.9811	0.9811	0.9811	0	0	0
GSInKB	0.7618	0.7618	0.7618	0.7607	0.7607	0.7607

Table 5.26: Results of GERBIL evaluation on CHEL using DBpedia and GeoNames as KB and content from Europeana.

	μ_C	$\mu_{C \text{ not NIL}}$	$F1_C$	P_C	R_C	$F1_{\text{NIL}}$	P_{NIL}	R_{NIL}
Micro	4.7374	4.8245	0.9468	0.9556	0.9383	0.0000	0.0000	0.0000
Macro	4.8708	4.8719	0.9446	0.9630	0.9445	0.9811	0.9811	0.9811

Table 5.27: Results for Hachey's on CHEL using DBpedia and GeoNames as KB and content from Europeana.

means of choosing a referent.

A further interesting result can be found when the value of $\mu_{C \text{ not NIL}}$ is examined a little deeper. While the arithmetic mean size of a candidate set is a little less than 5 entities, the median size was later found to be 1, meaning that most candidate sets, once again, were comprised of only a single entity. This is indicative of how little ambiguity was faced by HITS when choosing a referent. Outliers in the dataset such as "San Francisco", which had 322 candidate referents, and San Juan which had 290 candidate referents pulled the values for $\mu_{C \text{ not NIL}}$ up to an arguably misleading value. Table 5.28 provides a breakdown of counts for candidate sets of given sizes.

The scores for GERBIL's EE task and Hachey's P_{NIL}, R_{NIL} and $F1_{NIL}$ can largely be ignored for this evaluation. The vast difference between the micro and macro scores are due to the manner in which GERBIL internally handles division by zero errors. During the analysis of a document, if the values for tp, fp and fn are all zero, then GERBIL will assign a P and an R score of 1 for the task. Because the Europeana corpus contains no NIL annotations, CHEL achieves a perfect score for a document in the macro EE task provided it assigns a referent to every mention in said document. Given that most documents are comprised of only one entity, CHEL usually manages to do this. The result is an extremely high $F1_{macro}$ score.

In the micro task, because the result is based on the total values of tp, fp and fn across the whole corpus, the value for tp remains zero, but the value for fp increases for each instance of an erroneous NIL label. The result is that GERBIL assigns a score of zero for P_{micro} and R_{macro} in the EE task. In the interests

Candidate Set Size	Mention Count	Candidate Set Size	Mention Count	Candidate Set Size	Mention Count
1	802	11	1	23	1
2	218	12	12	24	17
3	55	14	7	32	3
4	50	15	1	41	1
5	23	16	2	44	1
6	17	17	4	52	1
7	9	18	15	54	37
8	2	19	1	99	2
9	11	21	3	290	1
10	4	22	3	322	1

Table 5.28: Breakdown of counts of candidates per mention. Values on the left are the size of the candidate set, values on the right are the number of mentions whose candidate set was the given size.

of maintaining a complete record of the experiment results, these values have been included in this thesis, but they are of little use beyond noting that Hachey's metrics, when implemented to handle division by zero errors in the same manner as GERBIL achieves the same values for P_{NIL}, R_{NIL} and $F1_{NIL}$ in the candidate sets.

What the high macro scores for the NIL annotations indicates is that CHEL usually returned at least one referent for each mention in an input document.

5.11 Summary

What can be seen from the depositions evaluation is that the challenges with linking entities in 17^{th} century Irish content stems from the available knowledge resources, rather than issues with the linking methods themselves. This is demonstrated by the results seen in Hachey's measures which illustrate how infrequently the correct referent for a mention is among the pools of candidate referents. The use of multiple KBs in conjunction with DBpedia yielded an appreciable increase in the quality of annotations applied to the depositions both in terms of Hachey's measures and the GERBIL evaluation.

While the use of multiple KBs resulted in both CHEL and REDEN performing significantly better than AGDISTIS, it is difficult to directly compare the relative performance of their linking methods. As mentioned in the previous paragraph, the application of Hachey's measures during the evaluation suggests that the linking methods are effective at identifying the correct referent when there is sufficient information available. Indeed the baseline evaluation in Section 5.7 shows that AGDISTIS and CHEL are on parity

when using the same source of information. Hence the improvement in performance comes from the ability of the EL system to link with respect to multiple KBs, rather than its ability to discern the correct referent from a pool of ambiguous candidates.

While the results of the ODNB and DIB evaluations were a little disappointing, these resources still produced an excellent improvement in CHEL's ability to abstain from annotating. Resources such as the ODNB and DIB ontologies simply do not exist in a format that is conducive to performing EL on specialised Irish collections. At least, certainly not in the context of a challenge that is as specific as the depositions.

Primary sources will be important for the construction of KB sources for collections such as the 1641 depositions. These sources are considerably more likely to contain references to entities that are of interest to scholars who work with these collections. However, formally structuring these resources into trustworthy KBs which list distinct entities is a task that will require input from historical scholars. Linking them to other available KB resources is an even greater task. While a computer scientist can expedite the resolution of entities in the sources through assistive technologies, ultimately this will be an undertaking that will require input from historians who are well versed in Ireland's past.

Testing CHEL on the Europeana corpus reveals some of the challenges faced by graph-based methods on certain content types. CHEL's extremely high $F1_C$ score showed that it performed excellently when finding the referent in the KB, but it occasionally failed to select the correct final referent. The challenge lies in the fact that there is usually only one entity per document, invalidating a basic assumption of a graph-based approach to EL. This is very interesting and illustrates the need for genuinely careful thought and consideration with regards to the nature of the content being annotated. If the context is small and the number of entities is few, then a graph-based measure may suffer.

For CH collections such as the depositions or Thibaudet, graph-based methods make sense. Their use is effectively imposed by the nature of the KBs that must be used to annotate collections such as these. Long form descriptions of entities are not always available for systems which consider entity context. Considering again the use of primary sources for 17^{th} century content, these resources are typically tabulated listings of information with little in the way of documentation that describes the entities in each record. Similarly, BnF which was used for Thibaudet, does not provide long form descriptions of entities which may be used to perform a comparison between the context of a mention and typical contexts for a candidate referent.

An implementation of CHEL is currently available for use on Github[5]. The service has been implemented as an e-service in FREME-NER [99] and will continue to be updated, extended and improved based on the findings of future work.

[5] `https://github.com/munnellg/CHEL`

Chapter 6

Conclusion

> *"There are some words, of course, that are better left unsaid – but not, I believe, the word uttered by my niece, a word which here means that the story is over."*
>
> — Lemony Snicket, *The End*

6.1 Research Question, Objectives and Achievements

The research question pursued in this thesis is:

> To what extent can entity linking be effectively performed on highly specialised text-based cultural heritage collections, such that an EL system can generate appropriate annotations for these textual corpora?

This question identified two specific properties that needed to be considered in the pursuit of an answer. The phrase "highly specialised" indicated that there was a need for specific domain knowledge on the part of an annotator in order to identify referents for entities. The term "appropriate" indicated that it was important for the system to not only apply a correct referent, but also to abstain from annotating where no such referent existed.

The goals of this thesis as stated in Chapter 1, and with their corresponding outcome as a result of this research are stated below.

There has been much research into EL since its inception. Reducing the vast array of existing EL services to a core set of assumptions, features and solutions is the first goal of this work.

The results of this goal are presented in the State of the Art in Chapter 2, which presents a comprehensive overview of standard EL methods along with a breakdown of how various systems approach EL according

to two general classes of similarity – local and global similarity – three broad types measure that can be used to approximate the values of these similarity functions – probability priors, contextual similarity, coherence.

Within these categories, a multitude of different EL approaches emerge which are diverse and creative in their approach to EL, but all of which essentially aim to maximise at least one of these similarity functions according to some combination of measures which can be grouped according to the three types of measure.

Identifying which of these core approaches yields the best EL performance on our target content type is the second.

This was investigated through the experiment discussed in Chapter 3. The investigation resulted in the creation of an annotated subset of the depositions that could be used to assess the quality of various EL systems as they were discovered.

The experiment used GERBIL as a platform for investigating methods of EL due to its implementation of standard, recognised evaluation metrics and comprehensive listing of EL services that could be included in an experiment. It was found, however, that some systems did not participate in the experiment correctly and ignored the supplied configuration information. Of the systems tested, only 2 were found to perform the test according to the supplied configuration. Of these two, a graph-based system called AGDISTIS which uses HITS to score a graph of candidate referents (a coherence approach) was deemed the more appropriate for further investigation.

This investigation also resulted in a quantitative measure of the severe lack of coverage for historical Irish entities in Wikipedia as a KB, leading to a search for more suitable KB sources. This process is discussed in Chapter 4.

As was stated in Chapter 1, an underlying assumption of this research, albeit not an actual goal, was to conduct research with an eye to the humanities scholars who work with the CH collections that comprised the test corpora. In addition to identifying Geohive and GeoNames as sources of entity URIs, ODNB and DIB were found to be two untapped sources of information about significant individuals in the history of Ireland. These were used to construct a new KB source that could be used in the context of Irish historical research and would be acceptable to the historical scholars who worked with the depositions.

Where no single suitable approach is found, this thesis will investigate alternative or unifying methods that are more suitable to the content type targeted by this research.

Based on the results of the conducted evaluation, AGDISTIS seemed like a promising candidate for annotating the depositions. However, while it did perform better overall in the evaluation, its performance was considerably lower than was desirable for the given corpus. However, further research yielded the discovery of REDEN, which applied an EL method based on multiple KB sources to deal with challenging

CH content. Like AGDISTIS, REDEN uses a graph-based approach: degree centrality, rather than HITS. However, REDEN used an extremely strict candidate selection process that could not deal with the noisy nature of the content of the depositions. Furthermore, it could only use one reference KB per EL problem.

The limitations of both AGDISTIS and REDEN led to the development of CHEL which approached EL as a problem to be solved using a coherence measure (HITS) and devised a method of indexing multiple KBs that would facilitate the use of several reference KBs during EL.

Experiments with CHEL as described in Chapter 5 demonstrated that this approach yielded a vast improvement in the quality of annotations applied to the depositions. The performance of the EL system jumped from an F1 score of 0.6083 to 0.7833 after the inclusion of multiple KB sources. This was also a drastic improvement over the performance of AGDISTIS which achieved an F1 score of 0.5979 in the baseline evaluation. Further experimentation showed that CHEL was competitive with REDEN when applied to the French literary texts that were the subject of REDEN's original research. CHEL also performed appreciably well when applied to content obtained from Europeana, a domain where the lack of context should have severely limited its performance, yet it still achieved an F1 score of 0.7607 in the standard evaluation.

6.1.1 Contributions

The first contribution of this work is manifest in the Cultural Heritage Entity Linker system. CHEL employs a novel indexing method during KB construction which uses GUIDs to resolve entities across multiple KB sources into a single unified KB that can be used for EL. The structure of the KB means that graph-based EL methods can avail of relationship information that is distributed across multiple KB sources. This method of structuring the KB facilitates the use of multiple reference KBs during EL, something which was not possible when using REDEN during experiments.

It was shown through the experiments in Chapter 5 that CHEL exhibits a dramatic improvement in performance on the 1641 depositions using three simultaneous KB sources. An added benefit of the use of reference KBs is that they also limit the scope of the EL system, giving content curators who wish to annotate a collection greater control over where their source URIs come from.

Additionally, CHEL's use of Monge-Elkan as a fuzzy retrieval method during candidate selection improves its ability to deal with noisy spelling in content such as the depositions. Even in a baseline evaluation using only DBpedia it can be seen by examining Hachey's measures that CHEL's ability to identify the correct referent during candidate selection exceeds that of AGDISTIS.

CHEL is freely available for download and use from Github[1].

The second contribution of this thesis is the range of resources that have been created as a result of this work.

[1] https://github.com/munnellg/CHEL

The ODNB and DIB ontologies[2] which have been created represent an important step forward for the application of EL technology to Irish collections. Setting aside obvious benefits (improved coverage and specificity of entities), a core advantage of these ontologies is that they are derived from resources created and controlled by historical experts.

Trust in the KB is an important consideration that was repeatedly raised and discussed with historical scholars who aided this research. Prior to this PhD, the closest analogue to the ODNB and DIB ontologies was the work of Christopher Yocum whose Irish-Gen ontology which focused primarily on medieval Ireland. The ODNB and DIB ontologies span a far greater time period, reaching from the 1^{st} century up to the 20^{th} century.

The ODNB and DIB ontologies will form a solid and trustworthy foundation for annotating new CH material as it is digitised by the historical scholars who collaborated on this work. It is expected in the future that it may even be possible to begin including information derived from primary sources into these ontologies, enabling them to grow as new information comes to light.

The 1641 depositions evaluation dataset[3] too provides a useful means of assessing the quality of EL systems that may be applied to similar CH collections. Prior to this PhD, no representative evaluation corpus existed which presented the challenges faced when applying EL to Irish cultural heritage material. The current Irish EL corpus is the first of its kind and highly challenging to work with. It is hoped that this corpus may be used to research further improvements in the quality of EL for CH.

The third contribution of this work is the linking method that was used to resolve the contents of ODNB and DIB to DBpedia. This linking method was shown to outperform DBpedia Spotlight when applied to the same task and is generic enough that it can be applied to other biographies that may need to be resolved to a KB such as DBpedia. It is worth stating again that this approach is essentially an EL method, and while the manner in which it is currently implemented is slow, with some optimisation it could be applied even more generally to standard EL problems.

Four research papers have been published, which describe the work that was conducted during this PhD.

> Munnelly, Gary; and Lawless, Séamus. “Investigating Entity Linking in Early English Legal Documents”. In the Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018. June 3rd–6th 2018. Fort Worth, TX, USA.

This paper presents the results of the experiment that was discussed in Chapter 3. It discusses the construction of the 1641 depositions corpus and the execution of the GERBIL experiment.

[2] https://github.com/munnellg/ODNB_DIB_Dataset
[3] https://github.com/munnellg/1641DepositionsCorpus

> Munnelly, Gary; and Lawless, Séamus. "Constructing a Knowledge Base for Entity Linking on Irish Cultural Heritage Collections". In the Proceedings of the 14th International Conference on Semantic Systems, SEMANTiCS 2018. September 11th–13th 2018. Vienna, Autria.

This paper presents the method by which the ODNB and DIB ontologies were constructed, including information about how the data was harvested, processed and structured into a semantic representation. The paper is intended to provide a guide for researchers facing similar challenges to that faced in this thesis, where suitable KB sources are sparse and new ones may need to be constructed.

> Munnelly, Gary; Pandit, Harshvardhan; and Lawless, Séamus. "Exploring Linked Data For The Automatic Enrichment of Historical Archives". In the Proceedings of the 3rd International Workshop on Semantic Web for Cultural heritage, SW4CH, held in conjunction with 15th Extended Semantic Web Conference, ESWC 2018. 3rd–7th June 2018. Heraklion, Crete, Greece.

This workshop paper discussed the efforts to convert the primary source historical content discussed in Chapter 4 into an KB source. It describes the rationale for attempting to perform this task and discusses some of the vocabularies that were considered for constructing the ontology.

> Munnelly, Gary, Caputo, Annalina, Lawless Séamus. "Linking Historical Sources to Established Knowledge Bases in Order to Inform Entity Linkers in Cultural Heritage." In the Proceedings of the 1st Workshop on Computational Methods in the Humanities, COMHUM 2018. June 4–5, 2018. Lausanne, Switzerland.

This paper discusses the linking method that was used to resolve entities in ODNB and DIB to DBpedia.

6.2 Future Work

The challenges faced in the pursuit of this PhD raised several questions that were deemed interesting for future research. Some of these are direct continuations of the work completed. Others are more tangential, but have been identified due to challenges encountered during the investigation of EL systems.

6.2.1 (Automatic) Creation of Resources for Cultural Heritage Research

From a practical perspective, one of the primary challenges faced during this PhD was the lack of available resources to apply the desired EL methods to the target corpora. This PhD has created a number of resources that will be useful to researchers attempting to solve similar challenges, but there is still room for the creation of further assets.

Future work will return to the problem of constructing a KB source from primary source material as described in Section 4.2. While a resolved KB was constructed from primary source material during this PhD, the accuracy of its content requires detailed verification on the part of a historian before it could be considered complete.

The 1641 depositions corpus that was created during this PhD facilitated the investigation and answering of the research question that was being pursued in this thesis. However, it would be extremely interesting to expand upon the current corpus with a greater number of documents and a more diverse semantic vocabulary. Due to the nature of the content of the collection, this requires the input and assistance of highly specialised annotators. The results of this PhD have generated interest from the historical community with regards to the kind of significant improvements that may be achieved with the investigated EL methods, and there are plans to expand on the research applied to the depositions using newly digitised CH material. This may involve the annotation of further historical documents.

Earlier research during this PhD looked at the possibility of automatically generating gold standards that could be used to perform evaluations. This line of research was originally inspired by the work of Junte Zhang who investigated user interaction with web portals based on server logs [122, 121].

Automatic gold standard creation has also been investigated in the context of EL by Ngomo et al. [82]. Ngomo's approach is limited with respect to the challenges faced in this PhD as it uses triples in an ontology to generate simple statements containing mentions of entities. This does not help to capture the linguistic anomalies seen in the depositions, and it introduces a similar problem to that seen with the Europeana task force corpus i.e. the corpus contains no NIL annotations and is a highly idealised representation of a corpus. None-the-less, the effort involved in constructing an EL corpus by manual means in this PhD makes the prospect of investigating methods of automatically generating test collections a highly appealing avenue of research.

6.2.2 Maintenance and Enhancement of chel

Arguably no piece of software is ever truly complete, and the same can be said of CHEL. Future work will continue to test and enhance the approach implemented with CHEL. In particular, as new KB sources are created their subsequent effects on CHEL's efficacy will be studied.

CHEL focused heavily on relationships between entities. This was a sensible decision given the information that was available throughout this researc. In many instances, we will know little beyond the fact that an entity existed and was mentioned in a given resource. Relationships between entities may be indicated by their co-occurrence in a series of documents.

Yet the use of ODNB and DIB shows that in some instances we may, in fact, have a long form description of individual entities where they are historically significant. There would be value in pursuing more research into the use of contextual features in the disambiguation of entities, although this will face

several challenges.

As has been noted several times, as we move backwards in time, the semantics and quality of language will change as we encounter variant spellings and usage of words in English. Further back again we will encounter Latin, at which point language models learned from ODNB and DIB are unlikely to be particularly helpful.

Even so, contextual features are undeniably important. Consider the linking method used to resolve ODNB and DIB to DBpedia which relied entirely on contextual features and achieved a very respectable F1 score over the data. There is value in pursuing this avenue of work.

It would also be useful to dig deeper into the relationships between entities. HITS proved to be an effective means of establishing global coherence between entities, but this was largely tested on Knowledge Bases whose structure was based on the DBpedia Ontology. If CIDOC-CRM were the sole ontology used, what would the impact on the performace of HITS be? Are the better graph measures for assessing the quality of relationships between entities?

Finally it is worth noting that CHEL does not consider the temporal nature of the collection on which it is working. For example, if it is known that the target corpus describes events in the 17^th century, then the linking process may be simplified by restricting the search for candidates exclusively to those who would have flourished at that time. Exploitation of features such as these are considered future work.

6.2.3 Investigating Trust Between Historians and Automatic Annotations

An avenue of research which arguably requires dedicated consideration is the subject of establishing trust between historians and the annotations produced by an automated process. There is a practical mistrust of automated processes that should be held by any good, sceptical researcher. It was often said during this PhD that no one should ever fully trust a machine.

There is also a more general concern that automated process such as EL rob scholars of control over their data, and pollute their collections with undesirable external sources. The choice to use resources such as ODNB and DIB were as much an attempt to remedy some of this concern as it was an effort to construct a more specialised CH KB. While larger European DH projects such as Europeana use KB sources like DBpedia and are now moving towards the use of Wikidata, the use of these resources in the academic, focused collections that were the subject of this research is generally considered inappropriate.

To provide a hypothetical scenario that may help to explain what is meant by this use of the word “inappropriate”, imagine a system which is designed to help academics write research papers. The system identifies passages, statements, claims etc. within the content of the paper as it is being written and supplies articles that support or refute claims. If this system were to use Wikipedia as its sole source of information, from a rigorous academic perspective its claims would be subject to question. If a researcher were to attempt to construct a bibliography based on these Wikipedia articles, it almost certainly would

not pass any form of peer review and may very well damage the reputation of the academic in question.

This is analogous to the problem faced when annotating primary source CH material with information obtained from a KB source. When a URI is applied to an entity in a document, the entity is also essentially being annotated with every claim, statement and error that is present in the KB source. For academics who devote large portions of their career to accurately extracting, organising and annotating data, this injection of information from a source that does not adhere to their standards of correctness and accuracy is not viewed as helpful. In fact, quite the contrary.

A further means by which concerns were allayed was through the provision of thorough provenance data at each stage of any automated process. This provenance data was not necessarily read by those who used the outputs of the linking process, but its existence provided reassurance that the source of an annotation could be checked, should it be subject to question.

This investigation of trust and appropriate application of automation in a historian's workflow is an important avenue for future research if computer scientists are to effectively apply themselves to problems in CH research.

6.2.4 Annotation of the Trinity College Digital Collections

Content in the Trinity College Digital Collections has previously been considered as corpora which could be used for EL research. While this line of work did not ultimately contribute to this PhD, it is still an interesting future application of the findings of this work.

Much content in the library is considerably less noisy and better organised than that of the depositions. Rather than working on primary source text which is difficult to parse, data in the library has been manually entered into the library system by a librarian. Essentially, this would mean that EL is being applied to well structured meta-data, no entirely dissimilar to that seen in the Europeana test corpus. Names of people, locations, important classification tags etc. have all been manually extracted and associated with content items. Hence the purpose of EL here would be to semantically uplift these collections to a more intelligent representation.

The temporal span of material in the Trinity College Digital Collections is extremely broad, spanning more than 1,000 years of Irish history. However the items in the collection, and the associated individuals therein are more noteworthy than those seen in a collection like the depositions, suggesting that annotating this collection may be an easier task than has already been tackled.

While the identification of KB sources for annotating the depositions was quite challenging, for the Trinity College Digital Collections resources such as the Getty Vocabularies[4] or the Library of Congress Linked Data Service[5] may provide the information needed, but this would require further investigation.

[4] http://vocab.getty.edu/
[5] http://id.loc.gov/

A preliminary investigation into the use of this collection produced the following publication:

> Munnelly, Gary; Koidl, Kevin; and Lawless, Séamus. "Exposing Ourselves: Displaying our Cultural Assets for Public Consumption". In the Proceedings of the 1st International Workshop on Accessing Cultural Heritage at Scale, ACHS, Co-Located with 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016. June 19th–23rd, 2016. Newark, NJ, USA.

6.2.5 Application of Annotations Supplied by Entity Linking Systems

Given that it is now possible to perform EL on the depositions, future work will also look at the possible uses of these semantic annotations for domains such as entity oriented search, personalisation, and more novel perspectives on CH collections (see Section 6.3.2).

Research into subjects such as entity oriented search are, by now, quite well developed [4, 51] and some works have looked specifically at how the output of EL systems can be used to enhance the quality of search in certain CH collections even when the quality of links provided are "crummy" [84]. Given that this research was essentially applied on the front lines of historical research, it would be extremely interesting to investigate how the annotations supplied by an EL system could actively assist researchers, rather than simply providing structure over a collection.

6.3 Complementary Parallel Work

During the course of research, opportunities arose to pursue questions that were related to this PhD, albeit not necessarily core. Rather than document these investigations under Future Work, the decision was made to describe them separately. These are subjects that will be pursued in relation to the work documented in this thesis, but were not originally envisioned as goals of this research.

6.3.1 Establishing Corpus Size Limits

One of the limitations of working with CH collections such as the depositions or KB sources such as ODNB and DIB is that their size can limit the types of resources that can be created from them. For example, the EL method developed in Chapter 4 requires the creation of a Word2Vec model. This cannot be done with a corpus that is the size of ODNB and DIB. The solution in this specific case would be to train the model on Wikipedia and then subsequently apply that model to ODNB and DIB. This will work because both resources are written using modern, contemporary English.

However, for a resource such as the depositions where the language is noisy and archaic, there may not be a large enough source of information to train a model that can be applied to the content of the collection. Previous work investigated corpus size limits for language modelling methods such as LDA [7], demonstrating the challenges faced when building language models over collections such as depositions.

Investigations into constructing language models for a low-resource collection like the depositions is an interesting challenge which will be considered in the future.

This work resulted in the following talk:

> Munnelly, Gary; O'Connor, Alexander; Edmond, Jennifer; and Lawless, Séamus. "Finding Meaning in the Chaos". Presented at the First Expert Workshop on Topic Models and Corpus Analysis Organized 2015. December 14th–15th 2015. Dublin, Ireland.

and the following publication:

> Hengchen, Simon; O'Connor, Alexander; Munnelly, Gary; and Edmond, Jennifer. "Comparing Topic Model Stability across Language and Size". Proceedings of the Japanese Association for Digital Humanities Conference, JADH 2016. September 12th–14th 2016. Tokyo, Japan.

Additionally, tools were produced to assess the stability of the language models being investigated. These tools are available on Github[6].

Given the clear challenges imposed by working in a low resourced domain such as the one tackled in this thesis, there is clear value in further pursuing this work.

6.3.2 Temporal Entity Random Indexing

Collaboration with an expert in temporal dynamics led to an investigation of the evolution of entities over time in a large news corpus [11]. Given the LDC New York Times corpus [98], the objective was to first recognise entities in the content of news articles and then measure how these entities evolved over time, assuming that an entity was mentioned several times over the span of the collection. Articles used for testing spanned the years 1987 to 2007.

This work provided an interesting alternative perspective on a problem termed "Entity Aspect Linking" by Federico Nanni [81]. Nanni performed Wikification, but rather than simply linking a mention to a Wikipedia article, he treated subsections of an article as references to a specific aspect of an individual's life. For example, Arnold Schwarzenegger was variously a body builder, an actor and a politician at various stages/aspects in his career. When linking an entity to the KB, Nanni attempted to identify not only the referent of a mention, but also what aspect of that entity was being discussed.

During the work with Caputo, a similar challenge was faced except that the aspects of entities were not known in advance. Instead the objective was to derive the different aspects of an entity by first

[6] https://github.com/munnellg/elltm

performing EL across the entire corpus, then applying Random Indexing [97] to determine a word space representation for each entity based on the context in which it occurred. Changes in aspects of an entity could be identified by recognising where two instances of an entity had been found in separate time periods, but the vector obtained through Random Indexing was significantly different.

This work produced the following publication:

> Caputo, Annalina; Munnelly, Gary; and Lawless, Séamus. "Temporal Entity Random Indexing". In the Book of Abstracts of the 28th Digital Humanities Conference, DH 2018. June 26th–29th 2018. Mexico City, Mexico.

The results of this work were promising, and are still the subject of ongoing work. While this research was performed on the LDC New York Times corpus, which is more contemporary than the depositions, the question remains as to whether or not it may be possible to examine the evolution of entities in historical corpora using methods similar to those that have been tested on the New York Times.

6.4 Concluding Remarks

This thesis has conducted an investigation into the application of EL techniques to highly specialised CH collections with a specific focus on 17^{th} century Irish material. This is a muddy problem that faces Irish historical scholars who continuously deal with demanding text-based collections in the course of their research. As a research topic, this subject was deeply challenging for reasons such as:

1. Excessive noise in the content of the collection
2. The need for highly specialised personal knowledge in order to interpret the collection
3. The extreme lack of available resources to reliably annotate the collection

Even by traditional investigative means, resources such as the 1641 depositions are difficult to work with. Yet the collection is reasonably typical of the kind of material handled by the historians who collaborated on this work. With the development of tools such as Transkribus [58] the process of automatically digitising these collections becomes faster and easier. However the need for tools to automatically organise these collections grows.

The EL system developed in this thesis has provided a means of automatically applying semantic labels to highly specialised Irish collections. CHEL has been rigorously tested in this context and has demonstrated that it is capable of reliably applying appropriate labels to the content of the depositions as indicated by the F1 scores obtained with GERBIL.

This thesis has also provided resources for testing the efficacy of tools that may be applied to challenging CH collections such as the depositions. The annotated 1641 corpus provides a means of testing new EL tools under highly challenging circumstances. This corpus is available on Github[7] and researchers are encouraged challenge their novel EL ideas with this collection of documents.

Furthermore, the ODNB and DIB ontologies[8] created during this research now form the basis of a reliable, trustworthy source of entities in this CH context. The issues with trust in the KB source were briefly discussed in Section 6.2.3. ODNB and DIB are two well reputed sources of information in the historical research community. It is hoped that the availability of these KB sources will encourage researchers to adopt and use EL tools to assist in the study of their historical collections.

Overall this thesis has opened the doors for a number of interesting avenues of research and has laid a solid foundation on which future work may be constructed.

[7] https://github.com/munnellg/1641DepositionsCorpus
[8] `https://github.com/munnellg/ODNB_DIB_Dataset`

Appendix A

Extracts from the Depositions

Conveying the nature of the depositions is perhaps best achieved by reproducing some examples of their contents. Below, the text from a number of depositions has been rendered to lend context to the challenge faced by this thesis. The documents have been reproduced in full with original spelling, deletions, margin notes, and historian modifications preserved. The samples are based on the original transcripts provided by historians.

Each document is printed twice: once in plain text format to facilitate ease of reading and once with TEI markup which demonstrates the information available to the computer when processing the text. There is a simple key for interpreting the plain text rendering based on that used by the 1641 Depositions Project website[1]:

- *italics* denote inline additions or modifications by historians.
- <chevrons> denote marginalia.
- {braces} denote damage to the original manuscript with a historian supplied substitution.
- [crochets] denote unclear text in the manuscript with a historian supplied substitution.
- ~~strike-throughs~~ denote text that was crossed out in the original manuscript.

A.1 The Deposition of Phillip Sergeant

A.1.1 Plain Text Rendering

587

Phillip Sergeant of the towne of Mountrath in the Queens County gent tennant and servant to Sir Charles Coote knight & baronet & overseer of his Lynnen and fustian workes ~~workes~~ sworne and examined

[1] http://1641.tcd.ie

deposeth and saith: That in the begining of the Rebellion in the Queens County that is to say about the xth of December 1641 The said Sir Charles Coote was at Mountrath aforesaid forceibly deprived robbed and dispoyled of fustians and lynnen cloth and cotten yarne worth in all 716 li. ix s. ster And that at the same tyme this deponent was deprived and dispojled of his service and imployment *&* of other his goodes to his now loss & damage of One hundred and fifty powndes sterling And the said Sir Charles was alsoe forceibly deprived & dispojled of his other goods chattells & estate in seuerall placs within the seuerall Countys of Leitrim the Kinges County Roscomon, & the Queens County of very greate value the particulers whereof he cannott expresse & had divers of his howses burned and spojled by the Rebells And further sayth That the Rebells that soe dispojled and robbed the said Sir Charles Coote *& the deponent* in the Queens County were and are fflorence ffitzpatrick <A> of Castletowne in the same County Esquire a Captain or great Comander of Rebells: whose wiffe (as this deponent hath bin credibly told by divers persons) sayd that she had but one hand, & hoped that shee should wash the same in the said Sir Charles Coots bloud, and the rebellious souldjers and servants of him the said fflorence whose names he knoweth not And this deponent further saith That one Mr John Nicholson of Mountrath aforesaid gent and ~~On Phillip Sargente~~ his wiffe in the begining of the Rebellion were perswaded and drawne by the said fflorence ffitzpatrick to come and bring all their goods with them (which were of good value) from Mountrath to Castletowne where the said fitzpatrick dwelled promissing that they & their goods should be there kept as saffe as his owne Liffe, And by those faire promisses the said ffitzpatrick getting possession both of their persons & goodes, they there behoulding daily cruelties & murthers vpon other English and belike suspecting the like to be exercised against themselues, ~~desired~~ fled away secretly ~~o n to~~ *to Mountrath* Mountrath to the howse of one Mortogh McAboy where they had much adoe to escape murthering that night And the next day the said fflorence ffitzpatrick being returned with his Rebellious souldjers from the seige of the fort of Leix to Mountrath the said Mris Nicholson earnestly begged vpon her knees to the said ffitzpatrick That she & her husband might haue his passe for their conveying to the fort of Leix aforesaid but he denyed to giue them any passe or convoy: notwithstanding he then and there received from her the scarfe *she* wore & he bade her shift for herself

588

for he would have noe more to doe with her; Wherevpon she left him And her hus{band} <A> & she hired for mony oneHarding of ~~M o~~ or nere Mountrath and oth{ers} to convoy her husband and her to the fort, whoe carrieing them from thence carried them into a wood telling them that was the more secure way: {But} when thy had them there they stabd Mr Nicholson & with a sword cleft her head downe to her shoulders & Left them both murthered there and stript o{f} their clothes As this deponent hath bin very credibly informed both by English a{nd} irish: and he beleeveth & partly knoweth their Informacion to be true, And heard it ~~very be~~ *alsoe* very credibly reported that those murtherers repented themsel{ves} that they had not ~~brought~~ ript Mris Nicholsons belly & taken out her grea{se} and fatt to haue made candles withall; And this deponent hath bin credibly to{ld} by divers both of the irish and English and beleeveth That some of {the} said fflorence ffitzpatricks souldjers by the direccions of him the said fflorence o{r his} <B> wiffe hanged o n a protestant by name William ffox at Castletowne aforesaid: {And} sent

away his wiffe in a Carr towards Mountrath with 2 children: But {by} the way they~~e~~ *Rebells* killd the children outright and wounded her & threw them all into an old saw pitt & cast timber vpon them. And there left her & the children vnder that tymber where she Lay Languishing for twoe day{es} vntill that An irish rebell passing by she called to him & desired him to giue her a little water to drinck: And he telling her he wold giue her her fill did instantly with a great stone knock out her braynes

<Dr J: Mr B>
Phillip Sarginte
Jur ~~vlt febr~~ *viij o Januar ij* 1643
Hen: Jones
Hen: Brereton

Q County o
Phillip Sergent Jur 8 Jan: 1643
10 dec
hand Intw [-]

+

A.1.2 TEI Rendering

```
<pb n="fol. 351r" pagenum="587" />
<p>
    587
    Phillip Sergeant of the towne of Mountrath in the Queens County
    gent tennant and servant to Sir Charles Coote knight &amp; baronet
    &amp; overseer of his Lynnen and fustian workes
    <del rend="strikethrough">workes</del> sworne and examined deposeth
    and saith: That in the begining of the Rebellion in the Queens
    County that is to say about the xth of December 1641 The said Sir
    Charles Coote was at Mountrath aforesaid forceibly deprived robbed
    and dispoyled of fustians and lynnen cloth and cotten yarne worth
    in all 716 li. ix s. ster And that at the same tyme this deponent
    was deprived and dispojled of his service and imployment
    <add place="inline">&amp;</add> of other his goodes to his now loss
    &amp; damage of One hundred and fifty powndes sterling And the said
    Sir Charles was alsoe forceibly deprived &amp; dispojled of his
    other goods chattells &amp; estate in seuerall placs within the
    seuerall Countys of Leitrim the Kinges County Roscomon, &amp; the
    Queens County of very greate value the particulers whereof he
    cannott expresse &amp; had divers of his howses burned and spojled
    by the Rebells And further sayth That the Rebells that soe
```

```
    dispojled and robbed the said Sir Charles Coote
    <add place="inline">&amp;amp; the deponent</add> in the Queens County
    were and are fflorence ffitzpatrick
    <note type="marginalia">A</note> of Castletowne in the same County
    Esquire a Captain or great Comander of Rebells: whose wiffe (as
    this deponent hath bin credibly told by divers persons) sayd that
    she had but one hand, &amp;amp; hoped that shee should wash the same in
    the said Sir Charles Coots bloud, and the rebellious souldjers and
    servants of him the said fflorence whose names he knoweth not And
    this deponent further saith That one Mr John Nicholson of Mountrath
    aforesaid gent and
    <del rend="strikethrough">On Phillip Sargente</del> his wiffe in
    the begining of the Rebellion were perswaded and drawne by the said
    fflorence ffitzpatrick to come and bring all their goods with them
    (which were of good value) from Mountrath to Castletowne where the
    said fitzpatrick dwelled promissing that they &amp;amp; their goods
    should be there kept as saffe as his owne Liffe, And by those faire
    promisses the said ffitzpatrick getting possession both of their
    persons &amp;amp; goodes, they there behoulding daily cruelties &amp;amp;
    murthers vpon other English and belike suspecting the like to be
    exercised against themselues,
    <del rend="strikethrough">desired</del> fled away secretly
    <del rend="strikethrough">o n to</del>
    <add place="inline">to Mountrath</add> Mountrath to the howse of
    one Mortogh McAboy where they had much adoe to escape murthering
    that night And the next day the said fflorence ffitzpatrick being
    returned with his Rebellious souldjers from the seige of the fort
    of Leix to Mountrath the said Mris Nicholson earnestly begged vpon
    her knees to the said ffitzpatrick That she &amp;amp; her husband might
    haue his passe for their conveying to the fort of Leix aforesaid
    but he denyed to giue them any passe or convoy: notwithstanding he
    then and there received from her the scarfe
    <add place="inline">she</add> wore &amp;amp; he bade her shift for
    herself
</p>
<pb n="fol. 351v" pagenum="588" />
<p>
    588
    for he would have noe more to doe with her; Wherevpon she left him
    And her hus
    <damage>
        <supplied>band</supplied>
    </damage>
```

```
<note type="marginalia">A</note> &amp;amp; she hired for mony
oneHarding of
<del rend="strikethrough">M o</del> or nere Mountrath and oth
<damage>
    <supplied>ers</supplied>
</damage> to convoy her husband and her to the fort, whoe carrieing
them from thence carried them into a wood telling them that was the
more secure way:
<damage>
    <supplied>But</supplied>
</damage> when thy had them there they stabd Mr Nicholson &amp;amp;
with a sword cleft her head downe to her shoulders &amp;amp; Left them
both murthered there and stript o
<damage>
    <supplied>f</supplied>
</damage> their clothes As this deponent hath bin very credibly
informed both by English a
<damage>
    <supplied>nd</supplied>
</damage> irish: and he beleeveth &amp;amp; partly knoweth their
Informacion to be true, And heard it
<del rend="strikethrough">very be</del>
<add place="inline">alsoe</add> very credibly reported that those
murtherers repented themsel
<damage>
    <supplied>ves</supplied>
</damage> that they had not
<del rend="strikethrough">brought</del> ript Mris Nicholsons belly
&amp;amp; taken out her grea
<damage>
    <supplied>se</supplied>
</damage> and fatt to haue made candles withall; And this deponent
hath bin credibly to
<damage>
    <supplied>ld</supplied>
</damage> by divers both of the irish and English and beleeveth
That some of
<damage>
    <supplied>the</supplied>
</damage> said fflorence ffitzpatricks souldjers by the direccions
of him the said fflorence o
<damage>
    <supplied>r his</supplied>
```

```
</damage>
<note type="marginalia">B</note> wiffe hanged
<del rend="strikethrough">o n</del> a protestant by name William
ffox at Castletowne aforesaid:
<damage>
    <supplied>And</supplied>
</damage> sent away his wiffe in a Carr towards Mountrath with 2
children: But
<damage>
    <supplied>by</supplied>
</damage> the way the
<del rend="strikethrough">y</del>
<add place="inline">Rebells</add> killd the children outright and
wounded her &amp; threw them all into an old saw pitt &amp; cast
timber vpon them. And there left her &amp; the children vnder that
tymber where she Lay Languishing for twoe day
<damage>
    <supplied>es</supplied>
</damage> vntill that An irish rebell passing by she called to him
&amp; desired him to giue her a little water to drinck: And he
telling her he wold giue her her fill did instantly with a great
stone knock out her braynes

<note type="marginalia">Dr J: Mr B</note>
Phillip Sarginte
Jur
<del rend="strikethrough">vlt febr</del>
<add place="inline">viij o Januar ij</add> 1643
Hen: Jones
Hen: Brereton

Q County o
Phillip Sergent Jur 8 Jan: 1643
10 dec
hand Intw
<unclear reason="illegible">
    <supplied>
        <del rend="strikethrough"> </del>
    </supplied>
</unclear>

+
```

```
</p>
<closer>
    <signed>
        <roleName type="Commissioner" />
        <name>Henry Brereton</name>
    </signed>
    <signed>
        <roleName type="Commissioner" />
        <name>Henry Jones</name>
    </signed>
</closer>
```

A.2 The Deposition of Joseph Joice

A.2.1 Plain Text Rendering

Joseph Joice of Kisnebrasney in the kings County gentleman sworne and examined deposeth and saith That after the Rebellion was begun in the County aforesaid vizt about the xxth of November 1641 This deponent for saffty fled to the Castle of knocknamease in the same County *then belonging to Lieutenant Peisley* where he was imployed to gouerne and looke to the kinges Armes and that howse where he and other souldjers manteined that Castle for ten months together though they were often sharply and strongly beseeged by the Rebells And in deed the defenders and other people in the Castle being many were soe bestraited and driven to such extreame want and misery that they were inforced and very glad to eate the flesh of horses doggs and Catts But haveing not enowgh of that nor any thing elce; about sevenscore men women and children were quite famished to death And such was the misery and want amongst the rest in the Castle <symbol> that one whoe was a Scochman pinched with extreame hunger privately in the night tyme opened the grave of a man that was buried within the liberties of the Castle and fed vpon the dead and buried mans flesh & one of the souldjers of the Castle partly espyring him tooke ayme & thincking him an enemy shott him through soe that he dyed, And that <symbol> Scochmans wife afterwards hanged to death her owne child and eate her flesh for want of meate: And yet god gaue such Liffe & incorragement to this deponent and the other souldjers that they killed many of the Assaylants whoe lay intrenched closse within musket shott of the Castle: Howbeit many women and children (that hungar forced out of the Castle to seeke for grasse and weedes to eate for want of food were taken and hanged or otherwise murthered by the Rebells. And further <A> sayth That about 4 dayes after he came first to the said Castle vizt vpon or about the 24th of November 1641 divers Rebells vizt those of the names and septs of the Carrolls, the Magheryes the McGillfoiles Coghlans & Moloyes and their souldjers and complices robbed all the brittish in the Cuntry thereabouts of their goods and amongst others they *or some of them* deprived Robbed ~~or otherwise~~ & dispoyled him this deponent of his howsholdstuff apparell Cattle horses sheepe Corne hay swyne and other goods and chattells worth six hundred pownds and expelled him

1180

expelled from from his howse and farme & burned his said howse & dispoyled him of both to his damage and losse of CC li. more *besid es his future losse of xx li. per annum* And the said <800 li.> Rebells alsoe ~~then and there stript this deponents wiffe and~~ att the same tyme depriued and dispoyled Richard Beard *gent* this deponents son in lawe at Ballendowne in the Kings County ~~gent~~ of all his howshold goods Corne Cattle Cowes sheepe horses and other goods and burned his howses all downe to the ground to his losse of six hundred pownds and aboue *besids 20 li . per annum by his farme in future 3 yeres proffits being lo st* [come ing] *to 60 li.* And the said Richard Beard being since dead his wife with three poore children is left a distressed widow without meanes of subsistence And the said Rebells at the same tyme alsoe robbed and dispoyled ffrancis domvill Inkeeper at Raghary in the Kings County of Cattle howsholdstuff & other goods worth 100 li. ster and burned all his howses to the ground, And left him with his wife and <symbol> 3 small Children are stript naked without meanes of livelihood, And further saith That Jane Moore of Dublin heretofore widow and now the wife of him this deponent since the Rebellion began hath bin at seuerall times Robbed and pillaged by the Rebells of her goods and chattells with 100 li. And alsoe saith That the Rebells in the Kings County aforesaid vpon or about the said 24th of day of November 1641 did at Tomah in the Kings County by force and Armes deprive Robb and <s> dispojle one ffrancis Medop of Tomah aforesaid Esquire of howsehold stuffe Corne Cattle horses Mares sheepe ~~horses~~ provision plate money debts *apparell* and other his goodes and chattells and burned his Markett towne and howses of Kisnebrasney Consistinge of about 20 howses of his owne building and of a number of other howses built by his tennants to his losse of fiue thowsand pownds at least And expelled him and his wiff child and family from their habitacion lands and meanes worth 400 li. per annum whereof the deponent accompteth hee hath Lost already 3 yeres proffitts comeing to 1200 li., and he is Like to loose the future proffits vntill a peace be established. And further saith that the said Mris Medop and her Child [-] *with* her husbands brother and all their family being about 20 persons were stript of their clothes and turned naked away in frost and snow ~~naked~~ soe as they were forced ~~themse~~ ~~*to wind or*~~ to wynd and cover themselues with ropes of straw to shelter *& keepe* them from starveing & in that posture went 30 myles on foote to Limrick:

<Dr J: H B>
Joseph Joyce
Jur vijo Jan: 1643
Hen: Jones
Hen: Brerton
1181

Henry Brereton
Henry Jones

A.2.2 TEI Rendering

```
<pb n="fol. 259r" />
<p>
    Joseph Joice of Kisnebrasney in the kings County gentleman sworne
    and examined deposeth and saith That after the Rebellion was begun
    in the County aforesaid vizt about the xxth of November 1641 This
    deponent for saffty fled to the Castle of knocknamease in the same
    County
    <add place="inline">then belonging to Lieutenant Peisley</add>
    where he was imployed to gouerne and looke to the kinges Armes and
    that howse where he and other souldjers manteined that Castle for
    ten months together though they were often sharply and strongly
    beseeged by the Rebells And in deed the defenders and other people
    in the Castle being many were soe bestraited and driven to such
    extreame want and misery that they were inforced and very glad to
    eate the flesh of horses doggs and Catts But haveing not enowgh of
    that nor any thing elce; about sevenscore men women and children
    were quite famished to death And such was the misery and want
    amongst the rest in the Castle
    <note type="marginalia">symbol</note> that one whoe was a Scochman
    pinched with extreame hunger privately in the night tyme opened the
    grave of a man that was buried within the liberties of the Castle
    and fed vpon the dead and buried mans flesh &amp;amp; one of the
    souldjers of the Castle partly espyring him tooke ayme &amp;amp;
    thincking him an enemy shott him through soe that he dyed, And that
    <note type="marginalia">symbol</note> Scochmans wife afterwards
    hanged to death her owne child and eate her flesh for want of
    meate: And yet god gaue such Liffe &amp;amp; incorragement to this
    deponent and the other souldjers that they killed many of the
    Assaylants whoe lay intrenched closse within musket shott of the
    Castle: Howbeit many women and children (that hungar forced out of
    the Castle to seeke for grasse and weedes to eate for want of food
    were taken and hanged or otherwise murthered by the Rebells. And
    further
    <note type="marginalia">A</note> sayth That about 4 dayes after he
    came first to the said Castle vizt vpon or about the 24th of
    November 1641 divers Rebells vizt those of the names and septs of
    the Carrolls, the Magheryes the McGillfoiles Coghlans &amp;amp; Moloyes
    and their souldjers and complices robbed all the brittish in the
    Cuntry thereabouts of their goods and amongst others they
    <add place="inline">or some of them</add> deprived Robbed
    <del rend="strikethrough">or otherwise</del>
    <add place="inline">&amp;amp;</add> dispoyled him this deponent of his
    howsholdstuff apparell Cattle horses sheepe Corne hay swyne and
```

```
    other goods and chattells worth six hundred pownds and expelled him
    1180
    </p>
<pb n="fol. 259v" />
<p>

    <add place="inline">expelled from</add> from his howse and farme
    &amp; burned his said howse &amp; dispoyled him of both to his
    damage and losse of CC li. more
    <add place="inline">besid es his future losse of xx li. per
    annum</add> And the said
    <note type="marginalia">800 li.</note> Rebells alsoe
    <del rend="strikethrough">then and there stript this deponents
    wiffe and</del> att the same tyme depriued and dispoyled Richard
    Beard
    <add place="inline">gent</add> this deponents son in lawe at
    Ballendowne in the Kings County
    <del rend="strikethrough">gent</del> of all his howshold goods
    Corne Cattle Cowes sheepe horses and other goods and burned his
    howses all downe to the ground to his losse of six hundred pownds
    and aboue
    <add place="inline">besids 20 li . per annum by his farme in future
    3 yeres proffits being lo st </add>
    <unclear reason="illegible">
        <supplied>
            <add place="inline">come ing</add>
        </supplied>
    </unclear>
    <add place="inline"> to 60 li.</add> And the said Richard Beard
    being since dead his wife with three poore children is left a
    distressed widow without meanes of subsistence And the said Rebells
    at the same tyme alsoe robbed and dispoyled ffrancis domvill
    Inkeeper at Raghary in the Kings County of Cattle howsholdstuff
    &amp; other goods worth 100 li. ster and burned all his howses to
    the ground, And left him with his wife and
    <note type="marginalia">symbol</note> 3 small Children are stript
    naked without meanes of livelihood, And further saith That Jane
    Moore of Dublin heretofore widow and now the wife of him this
    deponent since the Rebellion began hath bin at seuerall times
    Robbed and pillaged by the Rebells of her goods and chattells with
    100 li. And alsoe saith That the Rebells in the Kings County
    aforesaid vpon or about the said 24th of day of November 1641 did
    at Tomah in the Kings County by force and Armes deprive Robb and
```

```
    <note type="marginalia">s</note> dispojle one ffrancis Medop of
    Tomah aforesaid Esquire of howsehold stuffe Corne Cattle horses
    Mares sheepe
    <del rend="strikethrough">horses</del> provision plate money debts
    <add place="inline">apparell</add> and other his goodes and
    chattells and burned his Markett towne and howses of Kisnebrasney
    Consistinge of about 20 howses of his owne building and of a number
    of other howses built by his tennants to his losse of fiue thowsand
    pownds at least And expelled him and his wiff child and family from
    their habitacion lands and meanes worth 400 li. per annum whereof
    the deponent accompteth hee hath Lost already 3 yeres proffitts
    comeing to 1200 li., and he is Like to loose the future proffits
    vntill a peace be established. And further saith that the said Mris
    Medop and her Child
    <unclear reason="illegible">
        <supplied>
            <del rend="strikethrough"> </del>
        </supplied>
    </unclear>
    <add place="inline">with</add> her husbands brother and all their
    family being about 20 persons were stript of their clothes and
    turned naked away in frost and snow
    <del rend="strikethrough">naked</del> soe as they were forced
    <del rend="strikethrough">themse</del>
    <del rend="strikethrough">
        <add place="inline">to wind or</add>
    </del>
    <add place="inline"> </add> to wynd and cover themselues with ropes
    of straw to shelter
    <add place="inline">&amp; keepe</add> them from starveing &amp; in
    that posture went 30 myles on foote to Limrick:

    <note type="marginalia">Dr J: H B</note>
    Joseph Joyce
    Jur vijo Jan: 1643
    Hen: Jones
    Hen: Brerton
    1181

    </p>
</closer>
    <signed>
        <roleName type="Commissioner" />
```

```
        <name>Henry Brereton</name>
    </signed>
    <signed>
        <roleName type="Commissioner" />
        <name>Henry Jones</name>
    </signed>
</closer>
```

A.3 The Deposition of Ann Read

A.3.1 Plain Text Rendering

115

Ann Read the relict of Hilkiah Read Late of *Cancarrick in* the parish of Drumreligh in the Countie of Leitrim gentleman sworne & examined deposeth and sayth That about the xxiijth of october Last which was since the begining of the presente Rebellion ~~(her said husband being in England)~~ Her said husband & she were expelled from deprived robbed or otherwise dispoyled of their goodes & chattells being of the values following vizt Corne worth xlv li. howsholdgoodes & provition worth xl li. In due debts 40 li. In bookes & a lease 20 li. Cattle horses & hay worth Lxv li. In all amounting to CCx li. Besides the Rebells forcibly tooke away her husbands evidences and writings *of good value but* of the *iust* value whereof She cannott for the present give any estimate *which robbery & spoile w as don &* [ ] By and by <A> the meanes of donnell mcGowran *of the County of Ca u an* a Comander of Rebells & djvers of his ~~wicked &~~ rebellious *servants & wicked* crew whose names shee knows not, & by Donnell o Rely Henry ô Rely & Rose ô Rely this deponents Late servants And further sayth that Ellen <I> the wife of ~~d of~~ the said donnell o Rely haveing the nursing of a yong *male* sucking child of the deponents stripped ~~h im~~ *her* of ~~his~~ *new* clothes *as this deponent verily beleeveth* & brought h~~er~~ *e r* to this deponent whoe being stript of her meanes had not wherewith to releeve the child withall soe as hee ~~by could & famyne dyed~~, And another of her sonns called Stephen Read being about 6 yeres of age: was about the xth of ffebruary last 1641, in the howse of James Gray of the Cavan & goeing forth to play ~~*with*~~ him, there *then* gathered about him about six Irish children of that towne, whoe suddenly fell vpon him ~~su~~ & in such manner that some with stickes and some with stones burst & broke out his braynes putt out his eyes, & bruised his bodie extreamly: soe that he by theis wicked yong impes (which were none of them as shee is perswaded above viij yeres of age): ~~v not lo ng~~ *quickly* after djed & hadd beene killed owtright in the place had not an Englishwoman comen thither whoe tooke vp the dying child from them: saying to them she wondered they could fynd in their harts soe to deale with such a poore chyld: But they answered that they wold doe asmuch to her if they were able, as she & ~~one~~ Mris Gray afterwards told her this deponent &

<45 li.
40 li.

40 li.
20 li.
~~65 li.~~210>

115

116 And further sayth that John ô Rely, sonn to Edmund ô <A> Rely now of Clowater ~~late by~~ Mr Callams Castle is now by the Rebells made sherriff of the Countie of Cavan and that both the said John & Edmund ô Reley and one Phelim McGawrane gentleman Daniell McGawran gentleman & Charles McGawran all of the <B> Countie of Cavan: Richard Ashe of Lyssemanie in the parish of Drumlahen (whoe is gone from the Protestant Church to Masse, & was *Comissary* [~~C hancellor~~] of the bishops Cort) Phillip McHugh McShane ô Rely of Bellenegary in the parish of Castleterra, in the County of Cavan gent whoe now liveth at Mr Taylors howse in Balljhayes, ~~Phillip McMullmore in the County~~ <C:> Edmund ô Rely of Clowater in the Countie of Cavan gentleman & Garrott ô Rely of the parish of Dromlahin in the County of Cavan gent are & have beene in actuall Rebellion & have borne armes with and amongst the Rebells Robbed stripped & received the protestants goods taken from them, & have Comitted other outrages. & that Phillp mcMulmore ô Rely whoe hath beene formerly verie kynd to the robbed & spojled English & releeved them very much, doth now keepe and harbour the Rebellious souldjers, but thincketh he doth it for feare only: <And the reason why shee [-] conceaueth this to be *soe* ~~likly~~ is because shee heard som of the English which weare harboured by him ~~all~~ reporte <symbol> that the Rebells in Action *of rebellion* did call the said *Phillip* Realy an English churle <symbol> ackording to the Irish (badogh Sasonogh) because he would offer to releive any English; and threatned to burne his house:> And further sayth that this deponents husband comeing out of England to dublin & hearing of the Rebellion & being tould that this deponent & her children were robbed stript and ~~dyed~~ *dead* in a ditch: Hee being overcomen with greef & beleeve ~~th~~ing the same to be true fell into sicknes whereof he soone after dyed And this deponent haveing soe lost her husband 2 of her children & being robbed and stripped of all her meanes is

116

117 now by greefe and extreame want become the miserable object of pittie & hath not wherewith either to manteine herself or her *3* remayneing & surviveing children

Anne Read
Jurat 12o July 1642
Edw: Pigott
John Sterne

117

118

16 Leitrim

Ann Read Jur 12o July 1642
Cert fact

~~Intw~~
hand w
23 octo
63 210 li.

118

Edward Piggott
John Sterne

A.3.2 TEI Rendering

```
<pb n="fol. 39r" pagenum="115" />
<p>
    115
    Ann Read the relict of Hilkiah Read Late of
    <add place="inline">Cancarrick in </add> the parish of Drumreligh
    in the Countie of Leitrim gentleman sworne &amp; examined deposeth
    and sayth That about the xxiijth of october Last which was since
    the begining of the presente Rebellion
    <del rend="strikethrough">(her said husband being in England)</del>
    Her said husband &amp; she were expelled from deprived robbed or
    otherwise dispoyled of their goodes &amp; chattells being of the
    values following vizt Corne worth xlv li. howsholdgoodes &amp;
    provition worth xl li. In due debts 40 li. In bookes &amp; a lease
    20 li. Cattle horses &amp; hay worth Lxv li. In all amounting to
    CCx li. Besides the Rebells forcibly tooke away her husbands
    evidences and writings
    <add place="inline">of good value but</add> of the
    <add place="inline">iust</add> value whereof She cannott for the
    present give any estimate
    <add place="inline">which robbery &amp; spoile w as don &amp;
    </add>
    <unclear reason="illegible">
        <supplied>
            <add place="inline"> </add>
        </supplied>
    </unclear> By and by
    <note type="marginalia">A</note> the meanes of donnell mcGowran
    <add place="inline">of the County of Ca u an</add> a Comander of
```

```
Rebells &amp; djvers of his
<del rend="strikethrough">wicked &amp;</del> rebellious
<add place="inline">servants &amp; wicked</add> crew whose names
shee knows not, &amp; by Donnell o Rely Henry ô Rely &amp; Rose ô
Rely this deponents Late servants And further sayth that Ellen
<note type="marginalia">I</note> the wife of
<del rend="strikethrough">d of</del> the said donnell o Rely
haveing the nursing of a yong
<add place="inline">male</add> sucking child of the deponents
stripped
<del rend="strikethrough">h im</del>
<add place="inline">her</add> of
<del rend="strikethrough">his</del>
<add place="inline">new</add> clothes
<add place="inline">as this deponent verily beleeveth</add> &amp;
brought h
<del rend="strikethrough">er</del>
<add place="inline">e r</add> to this deponent whoe being stript
of her meanes had not wherewith to releeve the child withall soe as
hee
<del rend="doublestrikethrough">by could &amp; famyne dyed</del>,
And another of her sonns called Stephen Read being about 6 yeres of
age: was about the xth of ffebruary last 1641, in the howse of
James Gray of the Cavan &amp; goeing forth to play
<del rend="strikethrough">
    <add place="inline">with</add>
</del> him, there
<add place="inline">then</add> gathered about him about six Irish
children of that towne, whoe suddenly fell vpon him
<del rend="strikethrough">su</del> &amp; in such manner that some
with stickes and some with stones burst &amp; broke out his braynes
putt out his eyes, &amp; bruised his bodie extreamly: soe that he
by theis wicked yong impes (which were none of them as shee is
perswaded above viij yeres of age):
<del rend="strikethrough">v not lo ng</del>
<add place="inline">quickly</add> after djed &amp; hadd beene
killed owtright in the place had not an Englishwoman comen thither
whoe tooke vp the dying child from them: saying to them she
wondered they could fynd in their harts soe to deale with such a
poore chyld: But they answered that they wold doe asmuch to her if
they were able, as she &amp;
<del rend="strikethrough">
    <add place="inline">one</add>
```

```
    </del> Mris Gray afterwards told her this deponent &amp;

    <note type="marginalia">45 li.
        40 li.
    40 li.
    20 li.

        <del rend="doublestrikethrough">65 li.
        </del>210</note>
        115
    </p>
<pb n="fol. 39v" pagenum="116" />
<p>
    116
And further sayth that John ô Rely, sonn to Edmund ô
    <note type="marginalia">A</note> Rely now of Clowater
    <del rend="strikethrough">late by</del> Mr Callams Castle is now by
    the Rebells made sherriff of the Countie of Cavan and that both the
    said John &amp; Edmund ô Releу and one Phelim McGawrane gentleman
    Daniell McGawran gentleman &amp; Charles McGawran all of the
    <note type="marginalia">B</note> Countie of Cavan: Richard Ashe of
    Lyssemanie in the parish of Drumlahen (whoe is gone from the
    Protestant Church to Masse, &amp; was
    <add place="inline">Comissary</add>
    <unclear reason="illegible">
        <supplied>
            <del rend="strikethrough">C hancellor</del>
        </supplied>
    </unclear> of the bishops Cort) Phillip McHugh McShane ô Rely of
    Bellenegary in the parish of Castleterra, in the County of Cavan
    gent whoe now liveth at Mr Taylors howse in Balljhayes,
    <del rend="strikethrough">Phillip McMullmore in the County</del>
    <note type="marginalia">C:</note> Edmund ô Rely of Clowater in the
    Countie of Cavan gentleman &amp; Garrott ô Rely of the parish of
    Dromlahin in the County of Cavan gent are &amp; have beene in
    actuall Rebellion &amp; have borne armes with and amongst the
    Rebells Robbed stripped &amp; received the protestants goods taken
    from them, &amp; have Comitted other outrages. &amp; that Phillp
    mcMulmore ô Rely whoe hath beene formerly verie kynd to the robbed
    &amp; spojled English &amp; releeved them very much, doth now keepe
    and harbour the Rebellious souldjers, but thincketh he doth it for
    feare only:
    <note type="marginalia">And the reason why shee
```

```
        <unclear reason="illegible">
            <supplied>
                <del rend="strikethrough"> </del>
            </supplied>
        </unclear> conceaueth this to be
        <add place="inline"> soe</add>
        <del rend="strikethrough">likly</del> is because shee heard som
        of the English which weare harboured by him
        <del rend="strikethrough">all</del> reporte &lt;symbol</note>
        that the Rebells in Action
    <add place="inline">of rebellion</add> did call the said
    <add place="inline">Phillip</add> Realy an English churle
    <note type="marginalia">symbol</note> ackording to the Irish
    (badogh Sasonogh) because he would offer to releive any English;
    and threatned to burne his house:&gt; And further sayth that this
    deponents husband comeing out of England to dublin &amp; hearing of
    the Rebellion &amp; being tould that this deponent &amp; her
    children were robbed stript and
    <del rend="strikethrough">dyed</del>
    <add place="inline">dead</add> in a ditch: Hee being overcomen with
    greef &amp; beleeve
    <del rend="strikethrough">th</del>ing the same to be true fell into
    sicknes whereof he soone after dyed And this deponent haveing soe
    lost her husband 2 of her children &amp; being robbed and stripped
    of all her meanes is
    116
</p>
<pb n="fol. 40r" pagenum="117" />
<p>
    117
now by greefe and extreame want become the miserable object of pittie
&amp; hath not wherewith either to manteine herself or her
    <add place="inline">3</add> remayneing &amp; surviveing children
    Anne Read
    Jurat 12o July 1642
    Edw: Pigott
    John Sterne

    117
</p>
<pb n="fol. 40v" pagenum="118" />
<p>
    118
```

```
16 Leitrim
Ann Read Jur 12o July 1642
Cert fact

    <del rend="strikethrough">Intw</del>
    hand w
    23 octo
    63 210 li.

    118

</p>
<pb n="fol. 40r" />
<p>

</p>
<closer>
    <signed>
        <roleName type="Commissioner" />
        <name>Edward Piggott</name>
    </signed>
    <signed>
        <roleName type="Commissioner" />
        <name>John Sterne</name>
    </signed>
</closer>
```

Appendix B

Full GERBIL Comparative Evaluation

As was mentioned in Chapter 3, an evaluation of EL systems on the 1641 depositions was carried out using GERBIL as a platform for studying the performance of different EL systems. In total, 8 systems were tested but only 2 of these were actually found to be configured correctly for the experiment. The remaining 6 participants attempted to perform NER in spite of the fact that the experiment type was D2KB.

Even so, these results are indicative of the challenge posed by the depositions. It can be seen in the tables of results below, just how crippling the noisy language of the depositions is these state of the art EL systems. Therefore, the full table of results has been included as an indicator of the challenge presented by the depositions corpus.

It shall be emphasised however that these figures are by no means accurate due to the misconfiguration of the experiment and should not be used as the basis for any further research.

Table B.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 B.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 B.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 B.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**

Bibliography

[1] Eneko Agirre et al. “Matching Cultural Heritage items to Wikipedia.” In: *LREC*. 2012, pp. 1729–1735.

[2] Donald Harman Akenson. *The Irish diaspora: a primer*. Learning Links, 1996.

[3] Ayman Alhelbawy and Robert J. Gaizauskas. “Graph Ranking for Collective Named Entity Disambiguation.” In: *ACL (2)*. 2014, pp. 75–80.

[4] Krisztian Balog. “Entity Linking”. In: *Entity-Oriented Search*. Springer, 2018, pp. 147–188.

[5] Tim Berners-Lee, James Hendler, Ora Lassila, et al. *The semantic web*. 2001.

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009. Chap. 7.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3 (Jan 2003), pp. 993–1022.

[8] Kurt Bollacker et al. “Freebase: a collaboratively created graph database for structuring human knowledge”. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM. 2008, pp. 1247–1250.

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

[10] Razvan C Bunescu and Marius Pasca. “Using Encyclopedic Knowledge for Named entity Disambiguation.” In: *Eacl*. Vol. 6. 2006, pp. 9–16.

[11] Annalina Caputo, Gary Munnelly, and Séamus Lawless. “Temporal Entity Random Indexing.” In: *DH*. 2018, pp. 460–461.

[12] Richard Eckart de Castilho et al. “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)*. 2016, pp. 76–84.

[13] Diego Ceccarelli et al. “Dexter: an open source framework for entity linking”. In: *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*. ACM. 2013, pp. 17–20.

[14] Crm Cidoc. *The CIDOC Conceptual Reference Model*. 2003.

[15] Ronan Collobert et al. "Natural language processing (almost) from scratch". In: *Journal of machine learning research* 12.Aug (2011), pp. 2493–2537.

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

[17] Marguérite Corporaal and Jason King. "Irish global migration and memory: transnational perspectives of Ireland's Famine exodus". In: *Atlantic Studies* 11.3 (2014), pp. 301–320.

[18] Silviu Cucerzan. "Large-scale named entity disambiguation based on Wikipedia data". In: (2007).

[19] Joachim Daiber et al. "Improving efficiency and accuracy in multilingual entity extraction". In: *Proceedings of the 9th International Conference on Semantic Systems*. ACM. 2013, pp. 121–124.

[20] Jeffrey Dalton and Laura Dietz. "A Neighborhood Relevance Model for Entity Linking". In: *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*. OAIR '13. Paris, France, France: Le Centre De Hautes Etudes Intrrnationales D'Informatique Documentaire, 2013, pp. 149–156. ISBN: 978-2-905450-09-8.

[21] Christophe Debruyne et al. "Serving Ireland's Geospatial Information as Linked Data." In: *International Semantic Web Conference (Posters & Demos)*. 2016.

[22] Melvil Dewey. *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library*. Brick row book shop, Incorporated, 1876.

[23] Lee R. Dice. "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3 (1945), pp. 297–302. ISSN: 0012-9658. DOI: 10.2307/1932409.

[24] Mohnish Dubey et al. "Earl: Joint entity and relation linking for question answering over knowledge graphs". In: *International Semantic Web Conference*. Springer. 2018, pp. 108–126.

[25] Scott L. DuVall, Richard A. Kerber, and Alun Thomas. "Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators". In: *Journal of Biomedical Informatics* 43.1 (2010), pp. 24–30. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2009.08.004. URL: http://www.sciencedirect.com/science/article/pii/S1532046409001051.

[26] Jack Edmonds. "Paths, Trees, and Flowers". In: *Canadian Journal of mathematics* 17.3 (1965), pp. 449–467.

[27] I Efremova. "Mining social structures from genealogical data". PhD thesis. Technische Universiteit Eindhoven, 2016.

[28] Maud Ehrmann et al. "Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0." In: *LREC*. 2014, pp. 401–408.

[29] Samuel Fernando and Mark Stevenson. "Adapting Wikification to Cultural Heritage". In: *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. LaTeCH '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 101–106.

[30] Kate Fernie. "LoCloud: local cultural heritage online and in the cloud". In: *Uncommon Culture* 6.2 (2015), pp. 83–87.

[31] Paolo Ferragina and Ugo Scaiella. "Fast and accurate annotation of short texts with wikipedia pages". In: *IEEE software* 29.1 (2012), pp. 70–75.

[32] Paolo Ferragina and Ugo Scaiella. "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. 2010, pp. 1625–1628.

[33] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. "Incorporating non-local information into information extraction systems by gibbs sampling". In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2005, pp. 363–370.

[34] Suzanne Fischer. "Nota Bene: If You 'Discover' Something in an Archive, It's Not a Discovery". In: *The Atlantic* (June 2012). ISSN: 1072-7825.

[35] Antske Fokkens et al. "Biographynet: Extracting relations between people and events". In: *arXiv preprint arXiv:1801.07073* (2018).

[36] Linton C Freeman. "A set of measures of centrality based on betweenness". In: *Sociometry* (1977), pp. 35–41.

[37] Linton C. Freeman. "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3 (Jan. 1978), pp. 215–239. ISSN: 0378-8733.

[38] Francesca Frontini, Carmen Brando, and Jean-Gabriel Ganascia. "Semantic web based named entity linking for digital humanities and heritage texts". In: *First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*. 2015.

[39] Francesca Frontini, Carmen Brando Escobar, and Jean-Gabriel Ganascia. "REDEN ONLINE: Disambiguation, Linking and Visualisation of References in TEI Digital Editions". In: *Digital Humanities 2016*. 2016, http–dh2016.

[40] Octavian-Eugen Ganea et al. "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. 2016, pp. 927–938.

[41] GeoNames. *GeoNames*. 2018. URL: `http://geonames.org/` (visited on 10/19/2018).

[42] Ralph Grishman and Beth Sundheim. "Design of the MUC-6 evaluation". In: *Proceedings of the 6th conference on Message understanding*. Association for Computational Linguistics. 1995, pp. 1–11.

[43] Zhaochen Guo and Denilson Barbosa. “Robust entity linking via random walks”. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM. 2014, pp. 499–508.

[44] Ben Hachey, Will Radford, and James R Curran. “Graph-based named entity linking with wikipedia”. In: *International Conference on Web Information Systems Engineering*. Springer. 2011, pp. 213–226.

[45] Ben Hachey et al. “Evaluating entity linking with Wikipedia”. In: *Artificial intelligence* 194 (2013), pp. 130–150.

[46] Xianpei Han and Le Sun. “A generative entity-mention model for linking entities with knowledge base”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, pp. 945–954.

[47] Xianpei Han, Le Sun, and Jun Zhao. “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*. SIGIR ’11. Beijing, China: ACM, 2011, pp. 765–774. ISBN: 978-1-4503-0757-4. DOI: 10.1145/2009916.2010019. URL: http://doi.acm.org/10.1145/2009916.2010019.

[48] Bernhard Haslhofer et al. “Augmenting Europeana content with linked data resources”. In: *Proceedings of the 6th International Conference on Semantic Systems*. ACM. 2010, p. 40.

[49] Sebastian Hellmann et al. “Integrating NLP Using Linked Data”. In: *The Semantic Web – ISWC 2013*. Ed. by Harith Alani et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 98–113.

[50] Keld Helsgaun. “Solving the equality generalized traveling salesman problem using the Lin–Kernighan–Helsgaun Algorithm”. In: *Mathematical Programming Computation* 7.3 (2015), pp. 269–287.

[51] Annika Hinze et al. “Improving access to large-scale digital libraries through semantic-enhanced search and disambiguation”. In: *Proceedings of the 15th ACM/IEEE-CS Joint conference on digital libraries*. ACM. 2015, pp. 147–156.

[52] Johannes Hoffart et al. “Robust Disambiguation of Named Entities in Text”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, pp. 782–792. ISBN: 978-1-937284-11-4. URL: http://dl.acm.org/citation.cfm?id=2145432.2145521.

[53] Johannes Hoffart et al. “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”. In: *Artificial Intelligence* 194 (2013), pp. 28–61.

[54] Eduard Hovy et al. “OntoNotes: The 90\% Solution”. In: *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*. 2006.

[55] Nancy Ide and Jean Véronis. *Text encoding initiative: Background and contexts*. Vol. 29. Springer Science & Business Media, 1995.

[56] Antoine Isaac et al. *Comparative evaluation of semantic enrichments*. Tech. rep. Technical report, 2015.

[57] Paul Jaccard. "The Distribution of the Flora in the Alpine Zone.1". In: *New Phytologist* 11.2 (Feb. 1, 1912), pp. 37–50. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.1912.tb05611.x.

[58] Melina Jander. "Handwritten Text Recognition–Transkribus: A User Report". In: *The electronic Text Reuse Acquisition Project (eTRAP)* (2016).

[59] Sergio Jimenez et al. "Generalized Mongue-Elkan Method for Approximate Text String Comparison". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2009, pp. 559–570.

[60] Thorsten Joachims. "Making large-scale SVM learning practical". In: *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998, pp. 169–184.

[61] Grzegorz Kondrak. "N-gram similarity and distance". In: *International symposium on string processing and information retrieval*. Springer. 2005, pp. 115–126.

[62] Maria Koutraki, Farshad Bakhshandegan-Moghaddam, and Harald Sack. "Temporal Role Annotation for Named Entities". In: *Procedia Computer Science* 137 (2018). Proceedings of the 14th International Conference on Semantic Systems 10th – 13th of September 2018 Vienna, Austria, pp. 223–234. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.09.021. URL: http://www.sciencedirect.com/science/article/pii/S1877050918316260.

[63] Matt Kusner et al. "From Word Embeddings to Document Distances". In: *International Conference on Machine Learning*. 2015, pp. 957–966.

[64] Jens Lehmann et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195.

[65] V. I. Levenshtein. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10 (Feb. 1, 1966), p. 707.

[66] John Lodge and Mervyn Archdall. *The Peerage Of Ireland: Or, A Genealogical History Of The Present Nobility Of That Kingdom: With Engravings Of Their Paternal Coats Of Arms: Collected from Public Records, Authentic Manuscripts, Approved Historians, Well-attested Pedigrees and Personal Information*. Vol. 2. Moore, 1789.

[67] Hugo Manguinhas et al. "Exploring Comparative Evaluation of Semantic Enrichment Tools for Cultural Heritage Metadata". In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2016, pp. 266–278.

[68] Olena Medelyan, Ian H Witten, and David Milne. "Topic indexing with Wikipedia". In: *Proceedings of the AAAI WikiAI workshop*. Vol. 1. 2008, pp. 19–24.

[69] Pablo N Mendes et al. “DBpedia spotlight: shedding light on the web of documents”. In: *Proceedings of the 7th international conference on semantic systems*. ACM. 2011, pp. 1–8.

[70] Rada Mihalcea and Andras Csomai. “Wikify!: linking documents to encyclopedic knowledge”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 233–242.

[71] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[72] George A Miller. “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[73] David Milne and Ian H Witten. “An open-source toolkit for mining Wikipedia”. In: *Artificial Intelligence* 194 (2013), pp. 222–239.

[74] David Milne and Ian H. Witten. “Learning to Link with Wikipedia”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM ’08. New York, NY, USA: ACM, 2008, pp. 509–518. ISBN: 978-1-59593-991-3.

[75] Petar Mitankin, Stefan Gerdjikov, and Stoyan Mihov. “An Approach to Unsupervised Historical Text Normalisation”. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. DATeCH ’14. Madrid, Spain: ACM, 2014, pp. 29–34. ISBN: 978-1-4503-2588-2. DOI: 10.1145/2595188.2595191. URL: http://doi.acm.org/10.1145/2595188.2595191.

[76] Alvaro Monge and Charles Elkan. “The Field Matching Problem: Algorithms and Applications”. In: *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, pp. 267–270.

[77] Andrea Moro, Francesco Cecconi, and Roberto Navigli. “Multilingual Word Sense Disambiguation and Entity Linking for Everybody.” In: *International Semantic Web Conference (Posters & Demos)*. 2014, pp. 25–28.

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

[79] Diego Moussallem et al. “MAG: A multilingual, knowledge-base agnostic and deterministic Entity Linking approach”. In: *Proceedings of the Knowledge Capture Conference*. ACM. 2017, p. 9.

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

[81] Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. “Entity-Aspect Linking: Providing Fine-Grained Semantics of Entities in Context”. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. JCDL ’18. Fort Worth, Texas, USA: ACM, 2018, pp. 49–58.

[82] Axel-Cyrille Ngonga Ngomo et al. "BENGAL: An Automatic Benchmark Generator for Entity Recognition and Linking". In: *Proceedings of the 11th International Conference on Natural Language Generation*. 2018, pp. 339–349.

[83] Fabian Odoni et al. "On the Importance of Drill-Down Analysis for Assessing Gold Standards and Named Entity Linking Performance". In: *Procedia Computer Science* (2018).

[84] Alex Olieman et al. "Good Applications for Crummy Entity Linkers?: The Case of Corpus Selection in Digital Humanities". In: *Proceedings of the 13th International Conference on Semantic Systems*. ACM. 2017, pp. 81–88.

[85] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, Nov. 1999.

[86] Helena Iles Papaioannou. *Actually, Yes, It *Is* a Discovery If You Find Something in an Archive That No One Knew Was There*. en-US. June 2012. URL: https://www.theatlantic.com/technology/archive/2012/06/actually-yes-it-is-a-discovery-if-you-find-something-in-an-archive-that-no-one-knew-was-there/258812/ (visited on 12/18/2018).

[87] Francesco Piccinno and Paolo Ferragina. "From TagME to WAT: A New Entity Annotator". In: *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*. ERD '14. New York, NY, USA: ACM, 2014, pp. 55–62. ISBN: 978-1-4503-3023-7.

[88] Jakub Piskorski and Roman Yangarber. "Information Extraction: Past, Present and Future". In: *Multi-source, Multilingual Information Extraction and Summarization*. Ed. by Thierry Poibeau et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 23–49. ISBN: 978-3-642-28569-1.

[89] Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[90] Jonathan Raphael Raiman and Olivier Michel Raiman. "DeepType: multilingual entity linking by neural type system evolution". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[91] Delip Rao, Paul McNamee, and Mark Dredze. "Entity Linking: Finding Extracted Entities in a Knowledge Base". In: *Multi-source, Multilingual Information Extraction and Summarization*. Ed. by Thierry Poibeau et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 93–115. ISBN: 978-3-642-28569-1. DOI: 10.1007/978-3-642-28569-1_5. URL: https://doi.org/10.1007/978-3-642-28569-1_5.

[92] Lev Ratinov and Dan Roth. "Design challenges and misconceptions in named entity recognition". In: *Proceedings of the thirteenth conference on computational natural language learning*. Association for Computational Linguistics. 2009, pp. 147–155.

[93] Lev Ratinov et al. "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. 2011, pp. 1375–1384.

[94] Radim Řehůřek and Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[95] Giuseppe Rizzo et al. “Making Sense of Microposts (# Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge.” In: *# MSM*. 2015, pp. 44–53.

[96] Harald Sack. “The journey is the reward-towards new paradigms in web search”. In: *International Conference on Business Information Systems*. Springer. 2015, pp. 15–26.

[97] Magnus Sahlgren. “An introduction to random indexing”. In: *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. 2005.

[98] Evan Sandhaus. *The New York Times Annotated Corpus LDC2008T19, DVD*. 2008.

[99] Felix Sasaki et al. “Introducing FREME: Deploying Linguistic Linked Data.” In: *MSW@ ESWC*. 2015, pp. 59–66.

[100] Marijn Paul Schraagen et al. “Aspects of record linkage”. PhD thesis. Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University, 2014.

[101] Wei Shen, Jianyong Wang, and Jiawei Han. “Entity linking with a knowledge base: Issues, techniques, and solutions”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460.

[102] Rainer Simon et al. “Linking early geospatial documents, one place at a time: annotation of geographic documents with Recogito”. In: *e-Perimetron* 10.2 (2015), pp. 49–59.

[103] Rainer Simon et al. “Semantically augmented annotations in digitized map collections”. In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. ACM. 2011, pp. 199–202.

[104] Micheál Ó Siochrú, David Brown, and Thomas Bartlett. “The Down Survey and the Cromwellian Land Settlement”. In: *The Cambridge History of Ireland*. Ed. by JaneEditor Ohlmeyer. Vol. 2. The Cambridge History of Ireland. Cambridge University Press, 2018, pp. 584–607. DOI: 10.1017/9781316338773.026.

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

[106] Christina M Steiner et al. “Evaluating a digital humanities research environment: the CULTURA approach”. In: *International Journal on Digital Libraries* 15.1 (2014), pp. 53–70.

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

[108] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: A Core of Semantic Knowledge". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. Banff, Alberta, Canada: ACM, 2007, pp. 697–706. ISBN: 978-1-59593-654-7.

[109] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. "Fast random walk with restart and its applications". In: *Sixth International Conference on Data Mining (ICDM'06)*. IEEE. 2006, pp. 613–622.

[110] John Unsworth. "What is humanities computing and what is not?" In: *Defining Digital Humanities*. Routledge, 2016, pp. 51–63.

[111] Ricardo Usbeck et al. "GERBIL: general entity annotator benchmarking framework". In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, pp. 1133–1143.

[112] Ricardo Usbeck et al. "AGDISTIS-graph-based disambiguation of named entities using linked data". In: *International Semantic Web Conference*. Springer. 2014, pp. 457–471.

[113] Seth Van Hooland et al. "Exploring entity recognition and disambiguation for cultural heritage collections". In: *Digital Scholarship in the Humanities* 30.2 (2015), pp. 262–279.

[114] Denny Vrandečić and Markus Krötzsch. "Wikidata: A Free Collaborative Knowledge Base". In: *Communications of the ACM* 57 (2014), pp. 78–85. URL: `http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext`.

[115] Jörg Waitelonis and Harald Sack. "Named Entity Linking in #Tweets with KEA." In: *# Microposts*. 2016, pp. 61–63.

[116] Max De Wilde. "Improving Retrieval of Historical Content with Entity Linking". en. In: *New Trends in Databases and Information Systems*. Ed. by Tadeusz Morzy, Patrick Valduriez, and Ladjel Bellatreche. Communications in Computer and Information Science. Springer International Publishing, Sept. 2015, pp. 498–504.

[117] William Winkler. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". In: *Proceedings of the Section on Survey Research Methods*. 1990, pp. 354–359.

[118] William E Winkler. "The state of record linkage and current research problems". In: *Statistical Research Division, US Census Bureau*. Citeseer. 1999.

[119] Ian Witten and David Milne. "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*. 2008, pp. 25–30.

[120] Mohamed Amir Yosef et al. "Aida: An online tool for accurate disambiguation of named entities in text and tables". In: *Proceedings of the VLDB Endowment* 4.12 (2011), pp. 1450–1453.

[121] Junte Zhang and Jaap Kamps. "A search log-based approach to evaluation". In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2010, pp. 248–260.

[122] Junte Zhang and Jaap Kamps. “Search log analysis of user stereotypes, information seeking behavior, and contextual evaluation”. In: *Proceedings of the third symposium on Information interaction in context*. ACM, 2010, pp. 245–254.

[123] Wei Zhang et al. “NUS-I2R: Learning a Combined System for Entity Linking.” In: *TAC*. 2010.

[124] Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. “Robust and Collective Entity Disambiguation through Semantic Embeddings”. en. In: ACM Press, 2016, pp. 425–434. ISBN: 978-1-4503-4069-4.