Linking historical sources to established knowledge bases in order to inform entity linkers in cultural heritage

Gary Munnelly, Annalina Caputo, Séamus Lawless
Adapt Centre
Trinity College Dublin
Ireland
{gary.munnelly, annalina.caputo, seamus.lawless}@adaptcentre.ie

Abstract

A problem for researchers applying Entity Linking techniques to niche Cultural Heritage collections is the availability of Knowledge Bases with adequate coverage for their domain. While it is possible to generate a specialised Knowledge Base from available resources, this can result in a collection which is semantically annotated, but remains separate from other collections due to the use of a unique vocabulary. This paper presents a linking scheme for mapping a newly created Knowledge Base of significant Irish people to DBpedia for the purposes of both enriching the new Knowledge Base, facilitating integration with other collections and enabling multi-Knowledge Base Entity Linking which has been the subject of some research. The method is described and evaluated, showing that achieves a high level of performance on a new Knowledge Base constructed from the Oxford Dictionary of National Biography and the Dictionary of Irish Biography.

1 Introduction

While Entity Linking (EL) has seen much development over the years (Bunescu and Pasca, 2006; Milne and Witten, 2008; Ratinov et al., 2011; Yosef et al., 2011; Usbeck et al., 2014; Waitelonis and Sack, 2016; Brando et al., 2016), it is hindered by several limitations when applied to Cultural Heritage (CH) collections. Most notable is a significant under-representation of entities in common Knowledge Bases (KB) such as DBpedia (Agirre et al., 2012; Van Hooland et al., 2015; Munnelly and Lawless, 2018b). Consequently, EL systems which are informed by such KBs are ill equipped for annotating niche CH collections such as those studied by historians.

A possible solution is to construct tailored KBs from resources used by scholars investigating CH material. Such KBs would presumably be more appropriate for annotating the kinds of specialised collections in question. However, taking this approach can hobble one of the greatest benefits of annotating a collection with semantic resources, namely the ability to integrate with other collections which are annotated using the same vocabulary.

This paper discusses a linking method which was developed while constructing a specialised KB for notable Irish historical figures. The approach is intended to identify corresponding entities in DBpedia for each entity in the new KB where such equivalents exist. This facilitates communication between collections annotated with the specialised KB and others annotated with more general KBs. Moreover EL with respect to multiple ontologies is made possible, provided the EL service in question supports such an operation. Research by Brando et al. (Brando et al., 2016) has shown that it is beneficial to EL in CH when a specialised KB can be integrated with a more general one, and an EL service can perform linking across both resources in unison.

In Section 2 this paper discusses related work and surrounding context which motivated the development of this method. Section 3 describes the method in question. An evaluation is carried out in Section 4 and Section 5 provides concluding remarks.

2 Related work

The overarching research related to this paper investigates methods of performing EL on primary source Irish historical archives. This research focuses on two entity types – people and locations.

Using DBpedia as a KB, previous work has

shown that only ~23% of entities in a manually annotated gold standard subset of the collection could be annotated with a corresponding entity URI (Munnelly and Lawless, 2018b). This illustrates that an overwhelming number of entities in the collection cannot be identified without using an alternative KB. Furthermore, the challenging nature of the language in the collection (the documents are rife with spelling inconsistencies) means that EL systems struggle to identify a correct referent even when the entity exists in the KB.

The challenges faced with regards to geographic features have been largely mediated using either GeoNames (GeoNames) or GeoHive (Debruyne et al., 2016) as KBs. Both of these linked data resources have significantly better coverage of Irish geography than DBpedia. They also have the added benefit of attempting to identify and link against their counterparts in the DBpedia ontology where possible. This means that a collection which is annotated with geonames or geohive entities can communicate, at least in part, with a collection that is annotated with DBpedia.

Identifying a suitable KB to represent people in the collection is more problematic. It is an unfortunate fact that most individuals do not matter enough to be documented in any commonly available KB.

Two resources used by historians in this domain are the Oxford Dictionary of National Biography (ODNB)[1] and the Dictionary of Irish Biography (DIB)[2]. Both 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 in DBpedia, or at the very least limit the scope of the linker's search to entities that are relevant to this geographic region.

The goal of this work is to connect entries in ODNB and DIB with their corresponding entries in DBpedia, such that a new KB built on these resources would be linked with their counterparts in a larger, more established semantic resource where such counterparts exist. This also helps to identify entities in ODNB and DIB which are not yet documented in DBpedia, showing where an EL system that is informed by a KB based on ODNB and DIB may be better equipped for linking in Irish historical archives.

3 Method

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[3]. 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 7th 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 7th 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.

For each biography entry in ODNB and DIB $b \in \mathscr{B}$, 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 $\mathscr{P}_b$ is returned. The best matching DBpedia referent $p_b^* \in \mathscr{P}_b$ for a given biography is the one that maximises the expression:

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

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

1. http://www.oxforddnb.com/
2. http://dib.cambridge.org/
3. http://lucene.apache.org/solr/

For a given candidate $p \in \mathscr{P}_b$, 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) (Kusner et al., 2015) as implemented in gensim (Řehůřek and Sojka, 2010). 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 (Mikolov et al., 2013) trained on a Wikipedia dump excluding redirects, disambiguation pages etc.

The name similarity function Φ is based on the Monge-Elkan Method (Monge and Elkan, 1996). The biography title b_{title} and name of a candidate p_{name} are lower-cased and tokenized. Stop words are removed yielding two sets of tokens $\mathscr{T}_b$ and $\mathscr{T}_p$. The sets are added to a bipartite graph with edge weights computed using Jaro-Winkler similarity (Winkler, 1990). An optimal mapping $\mathscr{T}_b \mapsto \mathscr{T}_p$ is found using Edmond's blossom algorithm (Edmonds, 1965) giving $\mathscr{W}$, the set of weighted edges which comprise the mapping. Name similarity is the generalised mean of the edge weights in $\mathscr{W}$ as described by Jimenez et al. (Jimenez et al., 2009) where $m = 2$ in this experiment:

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

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

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

Where α and β are tuning parameters chosen such that $\alpha + \beta = 1$.

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

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

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

4 Evaluation

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 retrieval 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. For the initial investigation, the BAT Framework (Cornolti et al., 2013) was used to assess performance[4]. The choice to use BAT instead of the more commonly employed GERBIL (Usbeck et al., 2015) at this point in the evaluation was for scrutability of the results.

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.

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. This is somewhat disappointing at it suggests that the number of entities gained from using ODNB and DIB as source KBs is not as high as may be desirable. However, one must still remember 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 retrieval 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. Even so, it was found that lending some small weight to the similarity between

4. https://github.com/marcocor/bat-framework

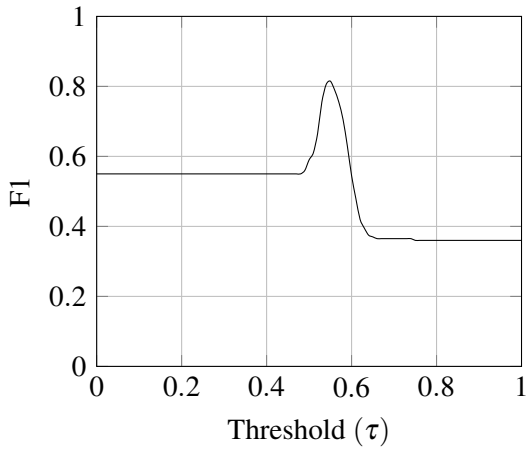

Figure 1: Change in performance on DIB with values of τ. Note that optimal threshold is seen when $\tau = 0.55$.

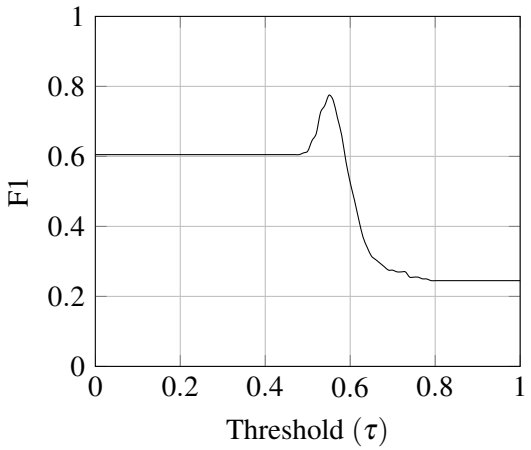

Figure 2: Change in performance on ODNB with values of τ. Note that optimal threshold is seen when $\tau = 0.55$.

surface forms yielded a slight increase in the $F1$ score for the linking method.

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 1 and 2, 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.

During the initial evaluation subject to the conditions above, this approach achieved an $F1$ score of 81.5% on DIB, but only 67.5% 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 an $F1$ score of 77.5%. 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	81.5
ODNB	0.55	67.5
ODNB (filtered)	0.55	77.5

Table 1: Summary of results

4.1 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. For this evaluation, the GERBIL benchmarking platform was used (Usbeck et al., 2015).

Annotator	Micro F1	Micro Precision	Micro Recall	Macro F1	Macro Precision	Macro Recall
Babelfy	0.5333	0.7304	0.4200	0.4200	0.4200	0.4200
DBpedia Spotlight	0.0099	0.5000	0.0050	0.0050	0.0050	0.0050
FOX	0.5112	0.5833	0.4550	0.4550	0.4550	0.4550
KEA	0.3437	0.3935	0.3050	0.3050	0.3050	0.3050
Munnelly	**0.8221**	**0.8241**	**0.8200**	**0.8200**	**0.8200**	**0.8200**
PBOH	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: GERBIL results for DIB. Munnelly (the method presented in this paper) clearly outperforms all other available services in the evaluation.

Annotator	Micro F1	Micro Precision	Micro Recall	Macro F1	Macro Precision	Macro Recall
Babelfy	0.6222	**0.8522**	0.4900	0.4900	0.4900	0.4900
DBpedia Spotlight	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FOX	0.4921	0.6667	0.3900	0.3900	0.3900	0.3900
KEA	0.5133	0.6259	0.4350	0.4350	0.4350	0.4350
Munnelly	**0.7700**	0.7700	**0.7700**	**0.7700**	**0.7700**	**0.7700**
PBOH	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 3: GERBIL results for ODNB. Munnelly (the method presented in this paper) clearly outperforms all other available services in the evaluation.

GERBIL is built on the BAT framework and so the results it produces were expected to be somewhat comparable with those presented in the previous section. However, in the interests of thoroughness a simple web interface to the biography linking method was set up so that GERBIL could directly benchmark the method.

The two gold standard collections – DIB and ODNB (filtered) – were converted into NIF documents (Hellmann et al., 2012). The surface form which named the subject of the biography was injected at the beginning of the article with some modification:

1. The names were originally in surname, forename order. This was reversed.
2. 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 EL systems. The generated surface form was marked as the only entity in the document. This clearly gives an advantage to EL systems which perform analysis on the context of a mention rather than relationships between entities. However, given the nature of the problem being tackled this is an appropriate bias.

GERBIL was configured to perform a D2KB evaluation, that is, the EL systems were provided with the surface form and the context of the mention. Their sole task was to identify a referent for the surface form.

At the time of the experiment, only 5 of the 17 EL services that are registered with GERBIL were available. These were Babelfy, DBpedia Spotlight, FOX, KEA, and PBOH (Moro et al., 2014; Mendes et al., 2011; Waitelonis and Sack, 2016; Speck and Ngomo, 2014; Ganea et al., 2016). The experiment was configured to run with these five services.

It should be noted that FOX is essentially AGDISTIS (Usbeck et al., 2014) with an entity recognition layer before the disambiguation phase. Given that this is a D2KB task, FOX can arguably be considered an evaluation of AGDISTIS with some caveats. Namely, FOX maintains its own deployment of AGDISTIS which is not necessarily in line with the most recent version, and the entity recognition stage in FOX is mandatory, meaning that even in the D2KB task it will attempt to spot entities. GERBIL compensates for this when computing the results of the evaluation, but it may still

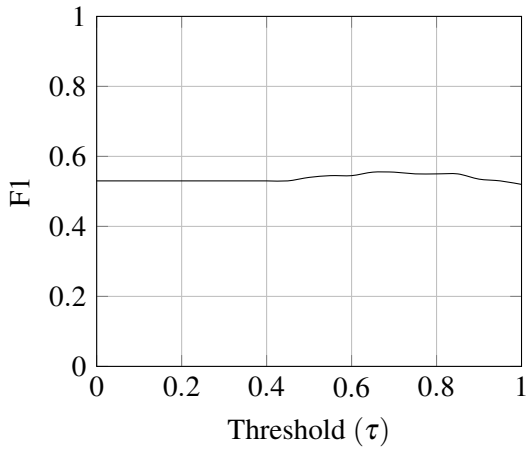

Figure 3: Change in performance of DBpedia Spotlight on DIB with values of τ. Note that the performance is much more consistent than earlier plots.

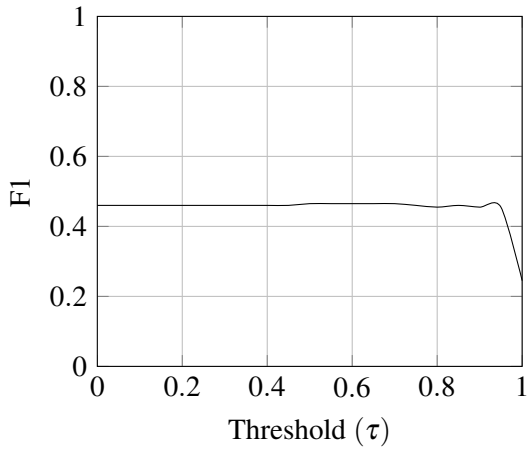

Figure 4: Change in performance of DBpedia Spotlight on ODNB with values of τ. Note that the performance is much more consistent than earlier plots.

lead to some skew in the final figures. Even so, the inclusion of FOX is helpful considering that the default AGDISTIS service was not online.

Under these conditions, GERBIL evaluated the EL systems. For brevity, only the precision, recall and F1 measures for the overall linking task are reported in tables 2 and 3. The full results of the evaluation on both DIB and ODNB are available online[5][6].

Both Macro and Micro $F1$ measures are reported. Macro and Micro take slightly different views of the collection. Macro treats each input document as an individual disambiguation problem, computing precision and recall for each document and then averaging the results across the whole collection. Micro treats the entire collection as one large disambiguation problem and computes precision and recall for all annotations in the gold standard. Given these definitions, we can expect that the results for Micro and Macro precision, recall and F1 measure will be roughly (if not exactly) equal for this specific evaluation, given that each document is comprised of only one entity. However, as previously mentioned, some services will attempt to perform Named Entity Recognition even when the specified task is D2KB. This can result in some disparity between the results of Micro and Macro evaluation.

The figures presented seem to confirm that the method described by this paper performs significantly better on this linking task than the other systems evaluated. However, it is difficult to understand precisely why this is the case since GERBIL does not provide access to the internal machinations of the evaluation. In particular, the performance of DBpedia Spotlight is unexpectedly low. Without knowing how GERBIL is calling this API endpoint, the reason for this seemingly complete failure cannot be determined.

Spotlight's performance is particularly suspicious as this is a task where it should perform reasonably well. It uses a language model to compare the context of a mention with descriptions of known entities, meaning it relies on contextual features to identify a referent; an approach which this experiment favours.

A specific evaluation using Spotlight's disambiguation API endpoint[7] was performed using a custom script using the content of each biography individually and the injected surface form as previously described. The responses from the server were dumped to a series of CSV files. As with the evaluation described in Section 4 the value of the confidence threshold for annotation was varied. Under these conditions, Spotlight performed considerably better with F1 scores reported by BAT between 0.25 and 0.465 for ODNB 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

5. http://gerbil.aksw.org/gerbil/experiment?id=201810190004
6. http://gerbil.aksw.org/gerbil/experiment?id=201810190005
7. http://model.dbpedia-spotlight.org/en/disambiguate

seen in Figures 3 and 4. It is notable they are much more stable than values shown in Figures 1 and 2.

This direct examination suggests Spotlight performs much better at this annotation task than the results of the GERBIL experiment indicate. This should not be considered as an attempt to undermine GERBIL, which is an important attempt at providing a consistent benchmark for the tumultuous challenge of evaluating EL systems. But it does strongly highlight the need for low level scrutability, reporting and configuration of the APIs being evaluated, which is a feature that GERBIL ostensibly lacks.

It is, nevertheless reassuring to see that the scores for this paper's method (designated "Munnelly" in the tables of results) conform to those values obtained by the earlier BAT investigation.

5 Conclusion

Given the task at hand, the method of linking presented in this paper 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 expected to become. This is extremely 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.

At the time of writing, the annotation task for linking ODNB, DIB and DBpedia has been completed and included in a custom KB (Munnelly and Lawless, 2018a). Ongoing work is investigating the usefulness of these links for improving the quality of EL on Irish CH datasets.

Acknowledgments

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

References

Agirre, Eneko, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando, and Mark Stevenson (2012). Matching cultural heritage items to wikipedia. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

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

Bunescu, Razvan C. and Marius Pasca (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *European Chapter of the Association for Computational Linguistics*, vol. 6, pages 9–16.

Cornolti, Marco, Paolo Ferragina, and Massimiliano Ciaramita (2013). A Framework for Benchmarking Entity-Annotation Systems. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 249–260. New York, NY, USA: ACM.

Debruyne, Christophe, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, and Declan O'Sullivan (2016). Serving Ireland's Geospatial Information as Linked Data. In *International Semantic Web Conference (Posters & Demos)*.

Edmonds, Jack (1965). Paths, Trees, and Flowers. *Canadian Journal of Mathematics*, 17(3):449–467.

Ganea, Octavian-Eugen, Marina Ganea, Aurelien Lucchi, Carsten Eickhoff, and Thomas Hofmann (2016). Probabilistic Bag-of-Hyperlinks Model for Entity Linking. In *Proceedings of the 25th International Conference on World Wide Web*, pages 927–938. International World Wide Web Conferences Steering Committee.

GeoNames (2018). Geonames. URL http://geonames.org/ (accessed 2018-10-19).

Hellmann, Sebastian, Jens Lehmann, and Sören Auer (2012). NIF: An Ontology-Based and Linked-Data-Aware NLP Interchange Format. *Working Draft*, page 252.

Jimenez, Sergio, Claudia Becerra, Alexander Gelbukh, and Fabio Gonzalez (2009). Generalized Monge-Elkan Method for Approximate Text String Comparison. In *International Conference on Intelligent Text*

Processing and Computational Linguistics, pages 559–570. Springer.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger (2015). From Word Embeddings to Document Distances. In *International Conference on Machine Learning*, pages 957–966.

Mendes, Pablo N., Max Jakob, Andrés García-Silva, and Christian Bizer (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. New York, NY, USA: ACM.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Monge, Alvaro and Charles Elkan (1996). The Field Matching Problem: Algorithms and Applications. In *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270.

Moro, Andrea, Francesco Cecconi, and Roberto Navigli (2014). Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In *International Semantic Web Conference (Posters & Demos)*, pages 25–28.

Munnelly, Gary and Séamus Lawless (2018a). Constructing a Knowledge Base for Entity Linking on Irish Cultural Heritage Collections. *Procedia Computer Science*, 137:199 – 210. Proceedings of the 14th International Conference on Semantic Systems 10th – 13th of September 2018 Vienna, Austria.

Munnelly, Gary and Séamus Lawless (2018b). Investigating Entity Linking in Early English Legal Documents. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*, pages 59–68. New York, NY, USA: ACM.

Ratinov, Lev, Dan Roth, Doug Downey, and Mike Anderson (2011). Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pages 1375–1384. Association for Computational Linguistics.

Řehůřek, Radim and Petr Sojka (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Speck, René and Axel-Cyrille Ngonga Ngomo (2014). Named entity recognition using FOX. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track*, pages 85–88. CEUR-WS.org. URL http://ceur-ws.org/Vol-1272/paper_70.pdf.

Usbeck, Ricardo, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both (2014). AGDISTIS – Graph-Based Disambiguation of Named Entities Using Linked Data. In *International Semantic Web Conference*, pages 457–471. Springer.

Usbeck, Ricardo, Michael Röder, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Brümmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. (2015). GERBIL: General Entity Annotator Benchmarking Framework. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1133–1143. New York, NY, USA: ACM.

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

Waitelonis, Jörg and Harald Sack (2016). Named Entity Linking in #Tweets with KEA. In *#Microposts—6th Workshop on Making Sense of Microposts*, pages 61–63. CEUR-WS.org. URL http://ceur-ws.org/Vol-1691/paper_14.pdf.

Winkler, William (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359.

Yosef, Mohamed Amir, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum (2011). Aida: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.