Linking Historical Sources to Established Knowledge Bases in Order to Inform Entity Linkers in Cultural Heritage

Gary Munnelly and **Annalina Caputo** and **Séamus Lawless**
Adapt Centre
Trinity College Dublin
firstname.lastname@adaptcentre.ie

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), it is hindered by several limitations when applied to Cultural Heritage (CH) collections. Most notable is a significant underrepresentation of entities in common Knowledge Bases (KB) such as DBpedia (Agirre et al., 2012; Van Hooland et al., 2015; Munnelly and Lawless, In Press). A possible solution is to construct more specific KBs from resources used by scholars investigating CH material.

The overarching research related to this paper investigates methods of performing EL on primary source Irish historical archives. 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 with a single entity usually being the focus of the text. 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 Hence they exhibit structural properties similar to those that originally made Wikipedia a useful KB for EL.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.

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. Hence 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.

[1] http://www.oxforddnb.com/
[2] http://dib.cambridge.org/

2 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. 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 names 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 the his name in DBpedia is given as "Dermot Bourke 7^{th} Earl of Mayo". Indexing anchor text loosely captures the equivalence of these two references.

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:

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

Where $\Psi(b, p)$ is computed as a linear combi-

nation 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) (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. Similarity is the negation of the normalised distance. 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 $\mathcal{T}_b$ and $\mathcal{T}_p$. The sets are added to a bipartite graph with edge weights computed using Jaro-Winkler similarity (Winkler, 1990). An optimal mapping $\mathcal{T}_b \mapsto \mathcal{T}_p$ is found using Edmond's blossom algorithm (Edmonds, 1965) giving $\mathcal{W}$, the set of weighted edges which comprise the mapping. Name similarity is the generalised mean of the edge weights in $\mathcal{W}$ as described by Jimenez et al. (Jimenez et al., 2009) where $m = 2$ in this experiment:

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

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.

3 Evaluation

The approach described is essentially an EL solution, hence the BAT Framework (Cornolti et al., 2013) was used for evaluation. Two ground truths were derived from a random sample of 200 biographies obtained from both DIB and ODNB (400 samples in total). Samples were manually assigned a DBpedia URI. 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.

A threshold similarity of $\tau = 0.55$ was found to give the best results when $\alpha = 0.1$ and $\beta = 0.9$, but the disparity in performance between the two collections is wide. Based on the evaluation, this approach achieves a 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. It is likely that the remaining errors with the approach are due to problems with document length normalisation as WMD is a measure of distance rather than relevance.

4 Conclusion

The approach documented seems promising given the scores achieved on DIB, but the wide variance in performance across the two collections indicates underlying problems. Future work will focus on making the approach more robust in an effort to eliminate this unreliability.

An objective of this work is to construct a KB which will be useful for EL on Irish historical archives. Future work will also investigate the applications of a KB, which is integrated with other semantic web resources, for EL, using the approach presented in this paper.

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

Eneko Agirre, Ander Barrena, Oier Lopez De Lacalle, Aitor Soroa, Samuel Fern, and Mark Steven-

son. 2012. *Matching Cultural Heritage items to Wikipedia*.

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

Razvan C Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Eacl*. volume 6, pages 9–16.

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

Jack Edmonds. 1965. Paths, trees, and flowers. *Canadian Journal of mathematics* 17(3):449–467.

Sergio Jimenez, 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*. Springer, pages 559–570.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*. pages 957–966.

Tomas Mikolov, 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.

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

Alvaro Monge 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.

Gary Munnelly and Séamus Lawless. In Press. Investigating entity linking in early english legal documents. In *Digital Libraries (JCDL), ACM/IEEE Joint Conference on*.

Lev Ratinov, 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-Volume 1*. Association for Computational Linguistics, pages 1375–1384.

Radim Řehůřek 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*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

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

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

Jörg Waitelonis and Harald Sack. 2016. Named entity linking in# tweets with kea. In *# Microposts*. pages 61–63.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. .

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