Matching Cultural Heritage items to Wikipedia

Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando, Mark Stevenson

IXA NLP Group, University of the Basque Country, Donostia, Basque Country,
{e.agirre,abarrena014,oier.lopezdelacalle,a.soroa}@ehu.es
Natural Language Processing Group, Sheffield University, Regent Court, 211 Portobello, Sheffield, UK
{s.fernando,r.m.stevenson}@sheffield.ac.uk

Abstract

Digitised Cultural Heritage (CH) items usually have short descriptions and lack rich contextual information. Wikipedia articles, on the contrary, include in-depth descriptions and links to related articles, which motivate the enrichment of CH items with information from Wikipedia. In this paper we explore the feasibility of finding matching articles in Wikipedia for a given Cultural Heritage item. We manually annotated a random sample of items from Europeana, and performed a qualitative and quantitative study of the issues and problems that arise, showing that each kind of CH item is different and needs a nuanced definition of what "matching article" means. In addition, we test a well-known wikification (aka entity linking) algorithm on the task. Our results indicate that a substantial number of items can be effectively linked to their corresponding Wikipedia article.

Keywords: Cultural Heritage, Corpus annotation, Wikification

1. Introduction

Current efforts for the digitisation of Cultural Heritage are providing common users with access to vast amount of materials. Europeana[1], for instance, is incorporating millions of digitised Cultural Heritage (CH) items from Europe's archives, museums, libraries and audio visual collections and providing access through a single portal. The main strength of Europeana lays in the vast number of items it contains. Sometimes, though, this quantity comes at the cost of a restricted amount of metadata, with many items having very short descriptions and a lack of rich contextual information. Wikipedia, in contrast, offers in-depth descriptions and links to related articles for many CH items, and is thus a natural target for automatic enrichment of CH items.

Enriching CH items with information from Wikipedia or other external resources is not novel. In (Haslhofer et al., 2010), for instance, the authors also acknowledge the interest of enriching CH items. They present the LEMMO framework, a tool to help users annotate Europeana items with external resources (i.e. Web pages, Dbpedia entries, etc.), thus extending Europeana items with user-contributed annotations.

In contrast to their work, our research aims to provide an evaluation of automatic annotation, and not only a description of an interface for manual annotation. We thus annotated a random sample of items, and performed a qualitative and quantitative study of the issues and problems that arise, showing that each kind of CH item is different and needs a nuanced definition of what "matching article" means. We also show that Wikipedia articles cover a substantial number of items.

Our research aims at finding Wikipedia articles that match the content of each target CH item. Note that this is more restrictive than finding Wikipedia articles that are related, as the matching article needs to describe the same CH object described in the target item. This problem is closely linked to Wikification, the process where a flat piece of text is enriched with links to the articles which are explicitly mentioned in the text. The process involves two inter-related steps: to choose which are the potential articles mentioned in the text, and to disambiguate them. For instance, assume that the famous Mona Lisa painting has been digitised and published as a CH item. In Wikipedia there are 11 articles which can be referred to as Mona Lisa[2], ranging from songs to a movie, and including actresses, singers and even a crater in Venus. In the first step of Wikification the algorithm would retrieve the 11 articles, and in the disambiguation step, the algorithm would select the painting[3]. Although a relatively recent concept, there is now a flurry of activity around this problem (Bunescu and Pasca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007; Milne and Witten, 2008; Han and Sun, 2011; Hoffart et al., 2011; Gottipati and Jiang, 2011; Ji and Grishman, 2011). We tested a well-known method (Milne and Witten, 2008) and our own in-house system on the task.

The paper is structured as follows. We begin by describing Europeana and the target collections. Section 3 presents the methodology for the manual annotation, followed by a Section analysing the annotated dataset. In Section 5 we describe the wikification systems used and the results when it is evaluated on our dataset. Finally, Section 7 draws the conclusions and outlines future work.

2. Europeana and the target collections

Europeana[4] is the prototype website of the European digital library. Europeana incorporates over 20 million digitised items from Europe's archives, museums, libraries and audio visual collections and provides access to them through a single portal. The need for personalised user services has been recognised from the early stages of Europeana's development. The items are supplied by over 1,500 institu-

[1] http://www.europeana.eu

[2] http://en.wikipedia.org/wiki/Mona_Lisa_(disambiguation)

[3] http://en.wikipedia.org/wiki/Mona_Lisa

[4] http://www.europeana.eu

```
<record>
  <dc:identifier>http://www.picturethepast.org.uk/frontend.php?keywords=Ref_No_increment;EQUALS;NCCW001197</dc:identifier>
  <europeana:uri>http://www.europeana.eu/resolve/record/09405/C052AA1727D9C258801CF676473953A0861A47C0</europeana:uri>
  <dc:title>The Major Oak</dc:title>
  <dc:source>Picture the Past OAI feed</dc:source>
  <dc:contributor>North East Midland Photographic Record</dc:contributor>
  <dc:description>The largest Oak tree in England, perhaps in the world, this famous tree has withstood lightning,
                  the drying-out of its roots and even a recent fire. The hollow tree has a circumference of 32 feet
                  and the spread of its branches makes a ring 260 feet round.</dc:description>
  <dcterms:isPartOf>Picture the Past</dcterms:isPartOf>
  <dc:language>EN-GB</dc:language>
  <dc:publisher>North East Midland Photographic Record</dc:publisher>
  <dc:subject>Robin_Hood</dc:subject>
  <dc:type>Image</dc:type>
  <dc:format>JPEG/IMAGE</dc:format>
  <europeana:provider>CultureGrid</europeana:provider>
  <europeana:hasObject>true</europeana:hasObject>
  <europeana:country>uk</europeana:country>
  <europeana:type>IMAGE</europeana:type>
  <europeana:language>en</europeana:language>
</record>
```

Figure 1: Example of an ESE record from Europeana.

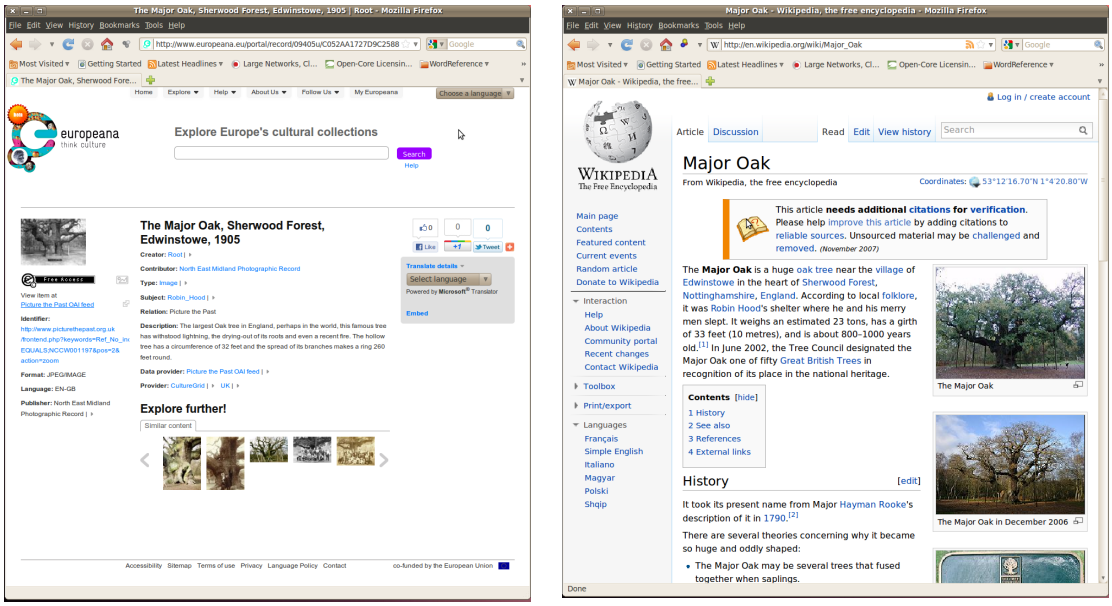

Figure 2: The Europeana item referring to a picture of the "The Major Oak" taken in 1905 (left), and the Wikipedia article on the same tree.

tions, including the British Library, the Louvre and other local museums, who have provided digitised items from their collections. We have focused on two of these collections: Culture Grid[5] and Scran[6].

Culture Grid (**Cgrid** for short) contains over one million items from 40 different UK collections including national and regional museums and libraries. The Scran collection is an online resource containing images and media from different museums, galleries and archives in Scotland.

The Europeana item records are associated with metadata which is extracted from the original collection though a process known as "ingestion". This paper uses a version of this metadata stored in a format known as Europeana Semantic Elements (ESE)[7]. Figure 1 shows an example of an ESE record describing a photograph of a well known tree, "The Major Oak". We focus on the `dc:title` and `dc:description` fields of the ESE records since the information they contain is relatively consistent (compared to other fields) and they generally contain enough text to work with. Figure 2 shows the item for the picture of "The Major Oak" as shown in the Europeana interface and the corresponding Wikipedia article referring to the same tree.

The combined collections contain approximately $858,000$ items, with $547,000$ items in Cgrid, and $310,800$ in Scran. Most of the items (99%) have a title ("dc:title"), which has 6 tokens on average, but only 68% have any description ("dc:description" field), with 27 words on average.

3. Methodology for a manually annotated dataset

We selected a random subset comprising 400 items from the Scran and Cgrid collections in Europeana. The items were then ordered according to the subcollections they came from, so the annotators had a relatively coherent set of

[5] `http://www.culturegrid.org.uk`

[6] `http://www.scran.ac.uk`

[7] `http://version1.europeana.eu/web/guest/technical-requirements`

Type	Count
Photographs	276
Coins and Artifacts	57
Books, booklets etc	24
Other	21
Paintings	14
Audio and Video	8
Total	400

Table 1: Types of Europeana items in the sample.

items, coming from a relatively small number of collections such as "The National Museum Record"[8], "The portables Antiquities Scheme"[9] or Scran. [10]. Table 1 shows the type of the items in the sample. The majority are photographs, but there are also other types such as paintings or antique coins.

The annotators were given the records with all the metadata (see Figure 1). They could also access the item as shown in the Europeana interface (see Figure 2) and they had to return the URL of a single English Wikipedia article (see Figure 2) matching the item, or NIL if they could not find any matching entry. The definition of a matching entry provided to the annotators was: "the Wikipedia article and the item must describe the same particular object. In the case of photographs, the article must be about the subject of the photograph, e.g a particular person or location." Note that this definition of matching tries to find equivalent items and articles, and thus does not consider other kinds of relations between item and Wikipedia article, such as for example linking an item to the article about "photography" because it's a photograph, or linking an item to the article of the author.

4. Analysis of the annotated dataset

The random subset of 400 items was independently tagged by two groups of annotators, one in Donostia and another in Sheffield, each one comprising three persons. As a result, the subset was annotated twice and two tags were obtained for each item. We chose one group's answers as gold standard, and used the other for calculating Inter Annotator Agreement (IAA) figures, as explained in Section 4.2.

According to the gold standard, 89 items were successfully linked to Wikipedia articles (22% of the sample). Given that the method for matching entries was very strict it came as a surprise that the annotators were able to identify a matching article for so many items. This result suggests that the task of matching Cultural Heritage elements to external resources such as Wikipedia can have a real impact in the richness of the descriptions for that 22% of the sample.

The remainder of this section describes the normalisation of URLs from Wikipedia to a canonical Wikipedia URL, followed by an analysis of agreement between annotators and qualitative analysis about what the annotators consider a "matching article" to be.

[8] http://viewfinder.english-heritage.org.uk/

[9] http://finds.org.uk/database/

[10] http://www.scran.ac.uk/

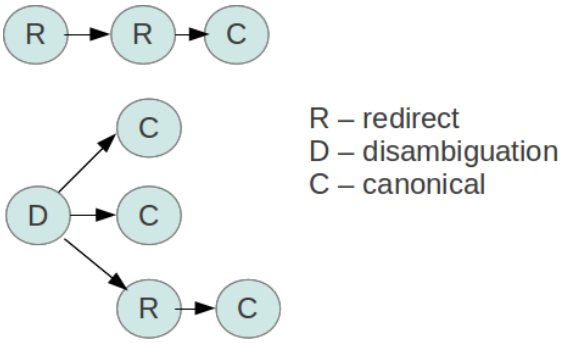

Figure 3: Normalisation flow.

4.1. Normalisation

Wikipedia articles are often accessed following so called redirect pages. For instances, the page "UK"[11] is a redirect page pointing to the Wikipedia article "United Kingdom"[12]. In such cases, we say that the redirect page "UK" resolves to the article "United Kingdom". Redirect pages fulfil many purposes like dealing with alternate names, plurals and closely related words.

When analysing the annotated items we found some discrepancies between annotators due to redirect pages: one annotator tagged the item with the "normal" article whereas other annotator used a redirect page resolving to the same article. We thus normalised the annotator results and resolved all redirect pages. In principle, it is enough to build a mapping among redirect pages and the articles they refer to. However, the process is further complicated due the fact that some redirects resolve to pages which are also redirects. Even worse, sometime redirect pages resolve to disambiguation pages (pages pointing to all possible meanings of a string) which can themself refer to other redirect pages.

Figure 3 shows two examples of the normalisation flow. The upper part of the figure shows a redirect resolving to another redirect which finally links to the desired article (the *canonical* article). The lower part shows an example of a disambiguation page referring to many pages; two of them are canonical pages but one page is a redirect which links to a canonical article.

The normalisation script thus builds a dependency tree between redirects, disambiguation pages and final articles. Then, it associates each article with a canonical link. Normal articles map to themselves; redirect pages map to the canonical page and disambiguation pages map to a set of possible canonical pages. Note that for our particular dataset no annotator chose a disambiguation page.

4.2. Inter Annotator Agreement

The overall Inter-Annotator Agreement (IAA) between the two tags available for each item are very high: 92.5% in the Cgrid collection and 80.0% in Scran (see Table 2). The agreement takes into account the items which were not associated with an article (i.e. tagged as NIL).

Given the high number of items with NIL, we also computed the IAA for items that were linked to an article for

[11] http://en.wikipedia.org/wiki/UK

[12] http://en.wikipedia.org/wiki/United_Kingdom

Match	Scran	Cgrid
Overall IAA	80.0%	92.5%
Agreement: Both NIL	126	165
Disagreement: One NIL	38	13
Agreement: Same article	34	20
Disagreement: Different articles	2	2

Table 2: Inter Annotator Agreement figures. The first row shows the percentage over all 400 items. The second and third rows show the numbers of items for which one of the annotations was NIL. The final two rows show the numbers of items where both annotators chose an article.

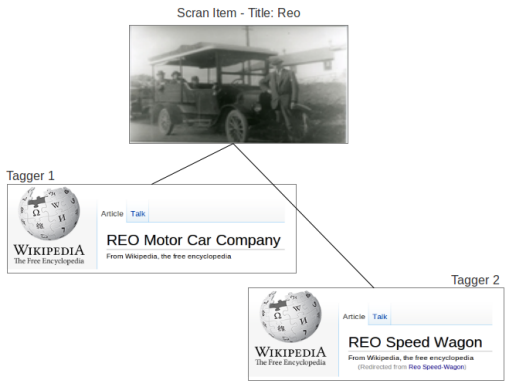

Figure 4: An example where the annotators did not agree. In this case both articles were acceptable.

both tags (22 items in Cgrid and 36 for Scran). The agreement for these items is even higher: 90.9% for Cgrid and 94.4% for Scran. We analysed the few cases where the taggers had both returned an article but would not agree, and found that in all cases both articles were acceptable, and very close in meaning. For instance, in one case an item about a light motor truck named 'Reo', manufactured by REO Motor Car Company was linked to the article about the truck model[13] by one tagger, and to the article about the company[14] by the other tagger (see Figure 4).

Most of the disagreements were due to one tagger not returning any article (NIL) and the other tagger choosing one article. We analysed these disagreements and in general the articles were relevant, and thus well linked. In a few cases, the article is not appropriate, although close. For instance, the item titled "Glyndebourne Opera Company present Le Comte Ory" containing one picture of a performance was linked to an article about an opera festival hold in Glyndebourne[15].

Overall the high IAA numbers (both overall and for items with no NILs) show that the annotation is reliable and that the task itself is well-defined.

[13] http://en.wikipedia.org/wiki/Reo_Speed-Wagon

[14] http://en.wikipedia.org/wiki/REO_Motor_Car_Company

[15] http://en.wikipedia.org/wiki/Glyndebourne_Festival_Opera

4.3. Qualitative analysis

Analysis of the annotations and the feedback received from the annotators suggested that the interpretation of "matching article" varied depending on the typology of the cultural item, as follows:

- If the item is a coin then the Wikipedia article judged as matching described the same kind of coin. For instance, the coin at http://www.europeana.eu/portal/record/09405v/0A9BB0DE9630F20665E36F10366069FDA3DAEA0D.html has no entry in Wikipedia, and therefore a NIL match would be returned. However, it is useful to consider cultural heritage items as instances of particular concepts. For instance, we can find an item about a particular antique coin like "Denarius, of Lucius Marcius Censorinus"[16]. As there is an article about this particular kind of coin[17], the annotators chose to link both.

- If the item is a picture of a particular location or person, that location or person was the subject of the matching Wikipedia article. For instance, for the item entitled "Hampton Court"[18] the matching article is http://en.wikipedia.org/wiki/Hampton_Court_Palace, but for the item "St. Leonards Church" [19] there was no matching article (even if the street is mentioned in the article on the town where it's located "Sunningwell"[20]). The same applies to people. For instance, the matching article for item "Albert Ball, Trent College"[21] s http://en.wikipedia.org/wiki/Albert_Ball. The same applies to organisations like soccer clubs. Note that pictures mentioning anonymous people (e.g. peasants) or locations can never be linked.

- If the title of the item mentions a person and a location annotators chose to focus on the person, as it's usually the focus of the picture. In the future, we would like to consider allowing double annotation.

- Many pictures are nearly 100 years old so there was sometimes a mismatch between the item of the picture and the more recent Wikipedia article.

5. Evaluating automatic systems

This section describes the evaluation of two automatic systems for linking Europeana items to Wikipedia when run on our dataset. Both systems take raw text as input, identify the possible anchors and link each to a Wikipedia article.

[16] http://www.europeana.eu/portal/record/00401/2FCF4C116A23D5F179CEE72DC9CAEE2A02721F79.html. Note that Europeana links change over time.

[17] http://en.wikipedia.org/wiki/Denarius_of_L._Censorinus

[18] http://www.europeana.eu/portal/record/09405r/7303A4578E3AE78F72EC75CB1F02DE47ECAFFE91.html

[19] http://www.europeana.eu/portal/record/09405o/F9C5A09A56B9C54DE0FCC9B53716716AAC751312.html

[20] http://en.wikipedia.org/wiki/Sunningwell#Parish_church

[21] http://www.europeana.eu/portal/record/09405u/03AD6F4A73D75F4BC5748E8AD2BA7096D45C7534.html

Wminer	acc	prec	recall	F1
Oracle1 Scran	0.240	0.206	0.650	0.313
Oracle1 Cgrid	0.240	0.122	0.724	0.209
Oracle2 Scran	0.895	0.672	0.650	0.661
Oracle2 Cgrid	0.960	0.750	0.724	0.737

Table 3: Oracle results for Wminer on Scran & Cgrid

Dict	acc	prec	recall	F1
Oracle1 Scran	0.200	0.200	0.667	0.308
Oracle1 Cgrid	0.115	0.111	0.759	0.193
Oracle2 Scran	0.900	0.667	0.667	0.667
Oracle2 Cgrid	0.965	0.759	0.759	0.759

Table 4: Oracle results for Dict on Scran & Cgrid

The experiments were carried out by providing each system with the text in the `dc:title` elements of the items.
The Wikipedia Miner toolkit (**Wminer** for short)[22] links entities found in a text to Wikipedia articles. The toolkit uses the method first presented in (Milne and Witten, 2008) which disambiguates terms by combining three features: the conditional probability of the article given the term (for example, the term "apple" is more likely to link to the article about the fruit than the one about the computer company), the probability of two terms appearing in Wikipedia as a collocation, and a vector-based similarity metric inspired by Normalized Google Distance (but using the links made to each Wikipedia article rather than Google's search results).
The second system uses a implementation similar to the dictionary method described in (Chang et al., 2010), which we refer to as **dict**. This method creates a dictionary containing information about the probability of a string matching a Wikipedia article. Each association between a string and article is scored by counting the number of times that the string appeared as the anchor text of an article's incoming hyperlinks. Note that such dictionaries can disambiguate any of the dictionary's keys directly by simply returning the highest-scoring article. We used the 2011 Wikipedia dump to construct the dictionary and are currently improving the linking algorithm to improve results using this approach.

5.1. Oracle results

In order to evaluate the automatic linking systems, we take the annotations of the first team as our gold standard (GS). We report separate results for the 200 items from Scran and the 200 items from Cgrid. We report accuracy (the ratio of items which get the same label as in the GS divided by the total number of items), precision (the ratio of items correctly linked to an article divided by the total number of items linked by the system), recall (the ratio of items correctly linked to an article divided by the total number of items of items linked to articles in the GS) and F1, the harmonic mean of precision and recall. Note that accuracy takes into account whether the system correctly assigns NIL, while the rest of measures only take into account items linked to articles (and thus discard items tagged as NIL).
Given the text in the title, a linking algorithm will return a set of articles, weighted according to the relevance assigned by the algorithm.
We first analysed whether the automatic linking algorithms are able to find matching articles, that is, whether the target matching article is contained among the articles they return. We are also interested in determining the upperbound in performance for a linking system which chose the correct matching article among the articles returned by the automatic systems. We set up two oracles:

- Oracle1: given a set of articles suggested by the wikifier for the item, choose the correct one (if available), otherwise return any article.
- Oracle2: if an item has no linked article in the GS (i.e. it was annotated as NIL) return NIL, regardless the output of the automatic system. Otherwise apply Oracle1, that is, given a set of articles suggested by the wikifier for the item, choose the correct one, if available

Tables 3 and 4 show the results for each oracle generated by the automatic systems. The accuracy of Oracle1 is very low (between 0.115 and 0.240, depending on the collection and system). The reason for this is that automatic systems suggest a matching article for most items while human annotators are much more selective and only link 22% of the items. The precision is also low for the same reason, as most of the articles returned by the systems were assigned NIL by the annotators. However, recall is high, ranging from 0.650 to 0.759. These figures are the upperbound for the recall of any automatic system built on the output of those wikifiers since the oracle selects all of the correct mappings which they return.
The Dict wikifier tends to return more articles than Wminer, in fact Dict always returns an article, and thus has a lower precision. The articles returned by Dict contain the correct article more often than Wminer as demonstrated by the higher recall figure on each of the collections.
Finally, the results for Oracle2 demonstrate the importance of choosing when to return NIL since a system which returns NIL with perfect accuracy (such as Oracle2) achieves high accuracy (between 0.895 and 0.965). The precision, recall and F-measures would also be high, with Dict generally outperforming Wminer by a small margin.
These results demonstrate that it is feasible to construct a system for automatically linking items to their matching Wikipedia entities based on the output of the Wminer and Dict methods. It is worth noting that we only use only used the text in the title for each item. The annotators mentioned that they used the information in the whole item, including the accompanying picture. This information was often an important factor in the annotators' decision to return NIL. In the future, we would like to explore whether performance could be improved by making use of information from other fields in the items.

[22] `http://wdm.cs.waikato.ac.nz/`

Wminer	acc	prec	recall	F1
Scran	0.190	0.153	0.483	0.233
Cgrid	0.205	0.081	0.483	0.139

Table 5: Results of applying the heuristics over the articles proposed by Wminer on Scran and Cgrid.

Dict	acc	prec	recall	F1
Scran	0.125	0.125	0.417	0.192
Cgrid	0.085	0.080	0.552	0.140

Table 6: Results of applying the heuristics over the articles proposed by Dict on Scran and Cgrid.

5.2. Article Selection Heuristics

We now explore a simple method for selecting the correct article from the set returned by the two methods. Information about the weights returned by each system is used alongside the start and end offsets of the words that were matched to the wikipedia article. In this preliminary study a simple algorithm based on the following set of heuristics is tested:

- Articles with high weights are preferred
- Articles matching longer strings are preferred
- Articles that match the start of the title are preferred

For Wminer the article with the highest weight is chosen first. In the case of ties the article with the longest matching string is chosen. If there is still a tie the article which matches closer to the start of the title is chosen.
The results for this heuristic are shown in Table 5. The main reason for the low accuracy and precision figures is that Wminer returns articles for items tagged as NIL. Recall is higher, 0.483 in both collections, showing that such a simple heuristic is able to select the correct article for nearly 50% of the items that have a corresponding Wikipedia article.
The Dict approach is somewhat different from Wminer since it returns a context-independent weight which is not comparable between articles and consequently the articles are set up in a different order. The articles with the longest match is chosen first and if there is a tie the one which matches closest to the start of the item title is selected. The weights returned by Dict are only used if there is still a tie. Results are shown in Table 6 which shows that the accuracy and precision figures are lower than those obtained using Wminer. The recall varies between the two collections and is lower for Scran than Cgrid.

5.3. Thresholding on weights

Given the over-generation of links for items, we analysed the effect of using the weight returned by Wminer to discard low scoring articles. The weights returned by Wminer ranged from 0.973 to 0.002, with an average of 0.480 on Scran and 0.477 on Cgrid. Ten thresholds lying between these values were selected. At each point we discard all articles with weights below the threshold. In this study we were interested in the ability to correctly identify cases

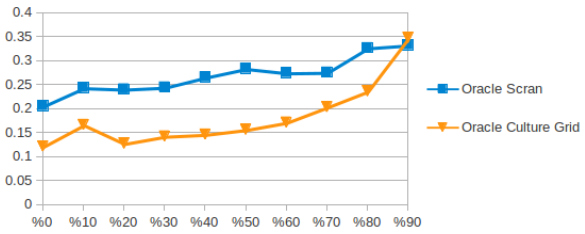

Figure 5: Precision of Oracle1 applied to Wminer weights filtered using various thresholds.

when there is not suitable article (i.e where the annotators selected NIL) as well as identifying the correct article and consequently Oracle1 is applied, i.e. for items linked to Wikipedia articles we choose the correct article if available among the choices returned by the system.
Figure 5 shows that Wikiminer weights are in principle useful to decide when to return NIL as precision raises for higher thresholds. However, the best precision that is achieved when the thresholds are applied is still well below the upperbound (precision for oracle2 is 67% and 75% for Scran and Cgrid respectively).
After the experiment we have seen that around 50% of the Wminer articles get weights in the lowest threshold band (under 10% of the maximum value). This explains why applying the heuristic used in Table 5 did not improve the results. It turns out that many correct articles have very low Wminer weights, and thus are discarded by the heuristic (but chosen by the Oracle1).

6. Conclusions and future work

In this paper we have performed an analysis of the issues that arise when Cultural Heritage items from Europeana are matched with Wikipedia articles. We have shown that up to 22% of items in Europeana can be matched with a counterpart in Wikipedia, a remarkable proportion when the vast number of items in Europeana is considered.
A well-know Wikification algorithm (Wikipedia miner) and an in-house method (Dict) were applied. It was found that up to 75.9% of the items matching a Wikipedia article could be linked automatically, given a perfect algorithm for choosing the correct one among the articles returned by the systems. A simple heuristic based on the weights returned by the systems, length and position in the title attains recall of 48.3% with Wminer and up to 55.2% with Dict (depending on the collection). The results are high for such a simple system, although the 75.9% upperbound shows that there is room for improvement. Note that we only used the text in the title, and an analysis of the text in the description could allow to find more and better matching articles.
We believe that the results reported in this paper are promising, and show potential for deploying a system which suggests Wikipedia articles for Europeana items. The main practical hurdle seems to devise a method which is able to decide when to abstain from returning an article, as there is a high ratio of items which do not have a corresponding

Wikipedia article and the automatic systems tend to always suggest articles. An initial study based on using the weights returned by Wminer showed promising results.

In future we plan to build a system which detects when to return NIL as well as improving techniques for selecting the correct article from those selected. We plan to achieve this by making use of more of the metadata associated with the item, and not only the title.

In addition, we also found that it could be useful to allow linking to subsections of Wikipedia articles, e.g in the case of streets or churches that are described inside the article of a town. For instance one of the Europeana items refers to Sunningwell parish church [23] and the article about Sunningwell includes a section on it[24].

Finally, in addition to identifying the best matching Wikipedia article it would also be interesting to identify related articles based on a fixed typology. For instance, in the case of an item showing the picture of a location, such as a monument or church, the system could return the article referring to the town in which the picture was taken.

Acknowledgments

The research leading to these results was carried out as part of the PATHS project (http://paths-project.eu) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082 and KNOW2 project (TIN2009-14715-C04-01). We want to thank the anonymous reviewers for their comments.

7. References

Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics.

Chang, A. X., Spitkovsky, V. I., Yeh, E., Agirre, E., and Manning, C. D. (2010). Stanford-ubc entity linking at tac-kbp. In *Proceedings of TAC 2010*, Gaithersburg, Maryland, USA.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716. ACL.

Gottipati, S. and Jiang, J. (2011). Linking entities to a knowledge base with query expansion. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 804–813, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Han, X. and Sun, L. (2011). 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*, pages 945–954, Portland, Oregon, USA. Association for Computational Linguistics.

Haslhofer, B., Roochi, E. M., Gay, M., and Simon, R. (2010). Augmenting europeana content with linked data resources. Proceedings of the 6th International Conference on Semantic Systems, pages 40:1–40:3, New York, NY, USA.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In Silva, M. J., Laender, A. H. F., Baeza-Yates, R. A., McGuinness, D. L., Olstad, B., Olsen, Ø. H., and Falcão, A. O., editors, *CIKM*, pages 233–242. ACM.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceeding of CIKM '08*, pages 509–518, New York, NY, USA.

[23] http://www.europeana.eu/portal/record/09405o/9215A3E5F9C4586ABB01D3EACFBA0B239AACDED4.html?query=Sunningwell

[24] http://en.wikipedia.org/wiki/Sunningwell#parish_church