Linking historical ship records to a newspaper archive

Andrea Bravo Balado, Victor de Boer, and Guus Schreiber

Department of Computer Science,
VU University Amsterdam,
Amsterdam, the Netherlands
a.c.bravobalado@student.vu.nl,{v.de.boer, guus.schreiber}@vu.nl

Abstract. Linking historical datasets and making them available on the Web has increasingly become a subject of research in the field of digital humanities. In this paper, we focus on discovering links between ships from a dataset of Dutch maritime events and a historical archive of newspaper articles. We apply a heuristic-based method for finding and filtering links between ship instances; subsequently, we use machine learning for article classification to be used for enhanced filtering in combination with domain features. We evaluate the resulting links, using manually annotated samples as gold standard. The resulting links are made available as Linked Open Data, thus enriching the original data.

Keywords: Text classification, machine learning, record linkage, entity linkage, historical research, digital humanities, digital history.

1 Introduction

Digital Humanities is a rising area of research at the intersection of disciplines in humanities with information technologies. This paper focuses on issues in *digital history*. Recently, many historical archives have been digitized. A key challenge in this area is increasing interoperability of heterogeneous datasets. Researchers in interdisciplinary settings are now focusing their efforts on linking historical datasets for enrichment and availability on the Web [11, 3].

In this paper we develop and evaluate a method for finding identity links between ships in two different datasets[1]. Our work is part of the Dutch Ships and Sailors (DSS) project. In this project maritime digital datasets have been made available as Linked Open Data [2]. The maritime history has been essential in the development of economic, social and cultural aspects of Dutch society. It has been well documented by shipping companies, governments, newspapers and other institutions.

Given the importance of maritime activity in every day life in the XIX and XX centuries, announcements on the departures and arrivals of ships or mentions of accidents or other events, can be found in historical newspapers, and

[1] This paper has an online appendix with technical details available at http://dx.doi.org/10.6084/m9.figshare.1189228

having these links available in the DSS data cloud would enrich the data, adding value for researchers. In this paper we describe how, using domain knowledge as well as machine learning text classification approaches, we are able to establish such links between different datasets for enrichment and help ease collaboration among historical researchers.

The method we use for linking involves a hybrid approach in which we use both domain-specific features to generate and filter candidate links, but also employ machine-learning techniques to improve the filtering process. We evaluate different combinations of these techniques, using manually annotated samples as gold standard and training set.

2 Related work

Many approaches for linking datasets can be found in the literature. For instance, the idea of using domain knowledge for entity linkage is not new. In [9], the task of entity linkage is focused on linking named entities extracted from unstructured text to the entities on a knowledge base. Although our approach is essentially the opposite, where we use a structured dataset (instead of a knowledge base) to find entities in the unstructured text and the domain is different, this area of research is related to ours. Even though the domain is different, the most similar to our goals is the research done in [13], where the authors present a system to disambiguate entity mentions in texts and link them to a knowledge base, the main difference being our use of a database in place of a knowledge base. Furthermore, in [5], in addition to domain knowledge, the authors take an information retrieval approach for linking entities. Moreover, finding and linking relevant newspaper articles has been done in [4] using a vector space model with a similarity function. The main difference with our research, besides the domain, is that the authors intend to link similar and current archives, while ours are essentially different and historical. Linking relevant newspaper articles from the Dutch National Library archives has been done in [6] and [8], albeit on a different domain, linking parliamentary and political debates with media outlets. Similarly, although their approach is by means of a semantic model and topic modeling, as well as using named entities for ranking, our experimental setup and evaluation procedure is based on [7].

Machine learning text classification has also been used for entity disambiguation and linkage. In [16], the authors experiment with text classification methods for literary study. Yu goes into detail about the importance of preprocessing and choice of classifiers, in which we have based some of our work. Moreover, in [15] the authors present information extraction as a classification problem to be solved using machine learning algorithms, such as Support Vector Machines (SVM) and Naive Bayes (NB), among others in order to extract information related to natural disasters from newspaper articles in Spanish.
Our work is part of recent efforts into linking historical datasets in the Netherlands, namely the works on linking datasets from German occupied Dutch society in [1] and historical census data in [12].

3 Approach

3.1 Datasets

We use two datasets which both contain descriptions of ship instances. The first dataset is the "Northern muster rolls databases" (in Dutch: *Noordelijke Monsterollen Databases*). This dataset contains official lists of crew members, known as "muster rolls", for ship companies in the three northern provinces of the Netherlands (Groningen, Friesland and Drenthe). The data was curated from mustering archives by historian J. Leinenga and covers the period 1803–1937. It was made available as Linked Open Data in the DSS project [2]. In this collection names of ships are not unique and may appear several times. In a preprocessing step we group ships which share (i) the same ship name, (i) the same last name of the captain, (iii) the same type of ship, and (iv) the same period (through a proximity relation of the respective record years).

The second dataset is the historical newspaper archive of the Dutch National Library (in Dutch: *Koninklijke Bibliotheek*. This data contains text and images of newspaper articles from 1618 to 1995 in the Dutch language. The newspaper archive is not limited to the maritime domain. The text of the articles has been digitized through OCR (Optical Character Recognition). The Dutch National Library indicates that the quality of the OCR text is not 100% reliable[2], due to common problems, such as old spelling, complex page layouts, difficult fonts, discoloration of the paper and fading of the ink. The data is available through a public website and API[3].

3.2 Evaluation

We perform a manual evaluation of the candidate links generated. Given the size of the dataset, it is unfeasible to manually assess every instance. Therefore, we have randomly selected a subset of 50 instances to be included in the evaluation for every experiment. Stasiu *et al.* [14] suggest that a sample of 50 instances is enough to extrapolate the evaluation to the rest of the dataset, according to their experiments for a similar problem.

For every candidate link we present the evaluator with full record information as well as the text of the linked article. For the experiments, which do not involve text-classification algorithms, the evaluation criterion is based on a 5-point Likert scale, ranging from strong disagreement (1) to strong agreement (5) on whether the newspaper text should be linked to a given ship. The 5-point Likert scale was requested by the domain expert and is also used for the calculation of mean and standard deviation. For precision, recall and F1 score calculations the Likert scale is transformed into binary scale, where values 1, 2 and 3 are considered non-relevant items (label 0) and values 4 and 5 are considered relevant (label 1). This is done to facilitate calculations. For the text-classification experiments a

[2] `http://kranten.delpher.nl/nl/pages/ocr`

[3] `http://www.delpher.nl/`

binary scale was used, where 0 indicates that there is no mention of a ship or ships in the text; 1 otherwise.

The evaluation was performed by historian and domain expert J. Leinenga (rater C), as well as by two co-authors of this paper (raters B and A). We measured the inter-rater agreement for each pair of raters by means of the weighted Cohen's Kappa coefficient (κ). Between raters A and B (0.76) as well as raters A and C (0.62) there is substantial agreement, whereas the degree of agreement between B and C is moderate (0.58).

For the assessment of our experiments, we have chosen to calculate the standard precision and F_1 scores, as well as an approximate recall. Precision is the fraction of candidate links that are evaluated as correct while the F_1 score is a weighted harmonic mean of precision and recall [10]. In order to calculate a recall score, given the difficulty to assess the actual number of relevant results in the newspaper archives, we propose an approximation, based on the estimated number of correct links retrieved by an algorithm divided by the estimated number of correct instances in the dataset, for which we take the baseline experiment, i.e. our most inclusive algorithm. The approximate recall is calculated by:

$$ApproximateRecall_x = \frac{\text{retrieved\_items}_\text{x} \cdot \text{precision}_\text{x}}{\text{retrieved\_items}_\text{baseline} \cdot \text{precision}_\text{baseline}} \tag{1}$$

4 Linking method

The linking method we deploy is a combination of processing steps. In the first step we generate a large candidate set of links; subsequent (alternative and/or consecutive) steps filter this large set with the aim to increase the link quality.

4.1 Baseline: Name of the ship and date restriction

We created a baseline by generating a large set of candidate links. We expect this baseline to have high recall and low precisions, as ship names are hardly unique: ships have typically names of females, of geographical locations, or of concepts such as friendship, hope and faith. The baseline was constructed by querying the API of the second dataset using the name of all ship instances from the first dataset. The baseline contains for every ship instance the first 100 links to newspaper articles provided by the API.

4.2 Domain feature filtering

For these processing steps our approach is to identify domain knowledge features that are suitable for distinguishing different ship instances and thus work well for filtering candidate links:

Filter 1a: Captain's last name In this filter we test a particular feature that domain experts have indicated can help on the disambiguation of candidate links from newspaper archives: the last name of the captain. This is mainly due to the fact that the captain of a ship, with few exceptions, is unlikely to change over time, unless the ship gets lost or destroyed. Note that this time, we are not performing queries on the full newspaper archive but only on the articles that contain at least a ship name and the publication year is within a range relevant for the ship instance at hand.

Filter 1b: Year restriction This filter helps to make sure that the year of publication of the candidate links from the newspaper archive is within the original muster-roll year interval for each ship instance. The rationale behind it is that ship instances that have already been found in historical records within given years, are less likely to be mentioned in newspapers published outside the typical lifespan of a ship, which is, according to domain experts, about 30 years.

Filter 2: Combining captain's last name and year restriction By combining the two filters 1a and 1b we expect to be able to boost precision, at the likely cost of recall.

4.3 Text classification

The processing steps using text classification are focused on exploring the structure of newspaper articles in order to train a classifier that would be able to predict labels for unseen data. The main difference with previous experiments is that the text classification process is not intended for generating or filtering links directly; they only return a value for whether or not this article describes a historical ship. We used two different classifiers: Naive Bayes and Support Vector Machine (SVM) with Sequential Minimal Optimisation (SMO).

We used the WEKA's[4] off-the-shelf supervised learning algorithms to train and evaluate a classifier model and then used the classifier to predict labels for the rest of the dataset. As training set we used the 200 labeled samples (121 positive and 79 negative instances), obtained during the manual evaluation of the previous steps (baseline, filters 1a, 1b, and 2). From these samples, we selected only the text of the corresponding newspaper article and we converted the 5-point Likert scale label, as explained before.

The next step was to choose and set multiple filters for data transformation. We make use of a multi-filter in order to apply all chosen filters at once. We chose the *string-to-word-vector* filter to represent the newspaper texts as feature vectors. For this experiment, we implemented a bag-of-words model, where the frequency of occurrence of each term is used as a feature, ignoring their order in the document [10]. We remove short words (< 3 characters) using the *remove by name* filter. Furthermore, we perform feature selection using a ranking based on the information gain metric. As recommended by WEKA documentation, the classifier is defined and evaluated but not yet trained. The evaluation is performed using both Naive Bayes and SMO classifiers and consists of a 10-fold cross validation using training data. Once the Naive Bayes and SMO classifiers

[4] http://www.cs.waikato.ac.nz/ml/weka/

were evaluated, these could be used for learning. The test set contains 413,663 instances. The algorithm returns a prediction; all new labels are imported into a new table for each classifier in the database to make it possible to associate labels to ship instances afterwards.

4.4 Combining domain feature filtering and classifier labels

The final step of our method involves combining techniques from the previous experiments. More specifically, we combined Filter 2 with the labels assigned by the classifiers as an additional feature. For these experiments, our method is to restrict the number of candidate links, as done during domain filtering. Similar to Filter 2, the last name of the captain and the restriction of the year of publication of the article are the domain features chosen for filtering. Additionally, we selected the label provided by the classifiers as positive examples in order to refine filtering of the candidate links, in the hope of improving our previous results. We test two variants of this combination, one with the Naive Bayes Classifier and the other with the SMO classifier.

5 Results

Table 1 summarises the results. The baseline contains 413,863 candidate links, corresponding to 5,078 ship instances. This baseline was used in all consecutive steps.

The first filter (captain's name) results in precision going up from 0.23 to 0.90, at the cost of a decrease in (approximate) recall from 1.00 to 0.40. The second filter (year restriction) gives a much smaller gain in precision (0.23 to 0.28) at a higher cost in recall (from 1.00 to 0.19). When we combine the two filters precision goes up to 0.96, at the cost of a further drop of recall to 0.13.

Table 1. Results for six (combinations of) processing steps: precision, approximate recall, F1 score, number of links retrieved, mean (λ) score, standard deviation (σ) of the score. * The two classifiers were evaluated using a binary scale instead of 5-point Likert scale and do not analyse candidate links but instead indicate whether an article is about ships.

Step	**Prec.**	**Approx. recall**	**F1**	**#Links**	(λ)	(σ)
Baseline	0.23	1.00	0.37	413,863	2.37	1.35
Filter 1a: Captain's last name	0.90	0.40	0.56	51,925	4.62	0.88
Filter 1b: Year restriction	0.28	0.19	0.23	79,113	2.58	1.58
Filter 2: 1a + 1b	0.96	0.13	0.23	16,037	4.80	0.49
Filter 2 + Naive Bayes	0.94	0.09	0.17	11,356	4.82	0.72
Filter 2 + SMO	0.94	0.10	0.18	12,215	4.84	0.51
Naive Bayes text classifier*	1.00	0.42	0.59	413,663	0.22	0.42
SMO text classifier*	1.00	0.45	0.63	413,663	0.30	0.46

Domain filtering combined with the two text classification methods resulted in a precision score of 0.94 and a similar approximate recall, of 0.09 and 0.10, respectively. These low recall scores affect the F1 scores, which are consequently low as well. This is mainly due to the restrictive nature of these algorithms, as evidenced by the number of retrieved links, being the lowest of all the experiments performed for this project.

The manual evaluation for the labels generated by the Naive Bayes and the SMO classifiers results in a precision of 1 and an approximate recall of 0.42 and 0.45, respectively. There were no false positives, thus the classifiers did not label relevant instances as irrelevant. Overall, the scores for the SMO classifier are somewhat better than the results of the Naive Bayes classifier.

6 Discussion

The baseline results show that it is indeed possible to retrieve a considerable amount of relevant links from a maritime historical dataset to historical newspaper articles. This despite the fact that OCR quality is imperfect and only a small part of the newspaper articles in the target dataset are about the maritime domain. Analysis of the baseline results also provided information for subsequent filtering steps.

When applying Filter 1a we found that the last name of the captain of a ship appears to be a good indicator for candidate link selection. This has also helped us gain more insight on the way ship instances are featured on the newspaper archives. By analysing the texts, we noticed that it is common to find the name of the ship along with the last name of the captain (either before or after), a port name and a date at the beginning of the sentence. For Filter 1b, on the one hand, we found that using the publication year and appearance in the muster-roll dataset do not yield successful results in terms of boosting precision. Even when the ship name appears in the text, only limiting the links to those of the years we have knowledge of is not enough for record linkage. However, we believe that this domain feature could be used either as a preprocessing step or in combination with more suitable domain features. Other features such as ship types were considered, but preliminary inspection showed that this feature does not appear in newspaper articles.

Precision can be further improved by combining domain features, as the results of Filter 2 show. We decided to concentrate on obtaining high precision scores because in general, historians prefer to have fewer but accurate links than the opposite. The results obtained by these experiments are satisfactory and have also given us an indication of the extent to which we can use domain knowledge for record linking.

The text-classification results by themselves (last two rows in Table 1) show that text classification in this domain is potentially a useful approach, regardless of the size of the training set. By analysing the feature vectors for both classifier models, we have found that the vectors consist mostly of port names (places) and female names, indicating that maritime activity terms are not mentioned on newspaper texts, at least not on common ship mentions. Also, our choices for data transformation and classifier configuration seem to be appropriate for the task at hand.

Our aim was to improve the link results by combining domain filtering and text classification. However, when we compare the scores of Filter 2 by itself to the scores of Filter 2 combined with either classifier we see that the latter have a negative impact on the results. Especially the loss in precision might be an artifact of the evaluation setup, especially given that the precision was already very high before this filter. Also, the restrictive nature of adding more features to the query causes loss of recall given that less links are retrieved. In retrospect, we think that a more sensible use of the classifier labels, e.g. as a filter before querying instead of it being part of the query, could have resulted in better scores. Still, more testing with bigger evaluation samples would be needed in order to decide whether combining both techniques is a suitable approach for this task.

We did not use the article titles in the training data for the classifiers. However, after updating the labels in the database, it was possible to see a distribution of article titles associated with each of the labels. Although there is some overlapping, it is noticeable that some newspaper sections seem more likely to be associated to ships than others, e.g. titles like “Advertentie” and “Familiebericht” (Eng: “Advertisement” and “Obituaries”) seem to be indicative of texts unrelated to ships while titles like “ZEETIJDINGEN.” and “Buitenl. Havens.” (Eng: “Sea messages” and “Foreign harbours”) are the opposite. These findings could be used as a feature for a different algorithm, e.g. for topic discovery. We believe it could also give an indication of the possible structures within the text articles given the distinctions between titles.

7 Conclusions

A total of 16,037 links resulting from Filter 2 have been used to enrich the main dataset as part of the DSS Linked Data cloud by De Boer et al. [2]. We chose the Filter 2 results because of the high precision, as this was high priority for the domain experts. These links provide new opportunities for analysis of the source material. Most found links are listings of arrivals or departures, including information about the destination. But other interesting examples of found links include links between ships and articles reporting on the sinking of that ship or the sale of the ship (including the price for which it was sold). These links are found to be interesting by maritime historical researchers. The data and example queries are available at `http://dutchshipsandsailors.nl/data`. Example queries include “return all newspaper articles between 1840 and 1850 for ships that sailed to Riga”.

In summary, we have successfully enriched a dataset of historical Dutch ships by linking its instances to corresponding mentions on newspaper archives. We explored different strategies for record linking and used both domain knowledge features and machine learning algorithms to link a dataset of historical Dutch ships with relevant entries in the newspaper archive of the National library of the Netherlands. Overall, we believe that our methods provide a limited but valuable set of links for historians and history enthusiasts that would have otherwise needed many hours of manual search and/or classification by experts. This is important for the accessibility of historical datasets on the Web as well as the preservation of these through time.

Acknowledgments. We would like to thank Jurjen Leinenga and our contacts at the Royal Library for their contributions. This work was supported by CLARIN-NL (http://www.clarin.nl) and by the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (http://esciencecenter.nl/).

References

1. de Boer, V., van Doornik, J., Buitinck, L., Marx, M., Veken, T., Ribbens, K.: Linking the kingdom: Enriched access to a historiographical text. In: Proceedings of the Seventh International Conference on Knowledge Capture. pp. 17–24. K-CAP '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2479832.2479849
2. de Boer, V., Leinenga, J., van Rossum, M., Hoekstra, R.: Dutch ships and sailors linked data cloud. In: Proceedings of the International Semantic Web Conference (ISWC 2014) (2014)
3. Boonstra, O., Breure, L., Doorn, P.: Past, present and future of historical information science. Historical Social Research / Historische Sozialforschung 29(2) (2004), http://www.gla.ac.uk/centres/hca/ahc/docs/pastpresentfuture.pdf
4. Bron, M., Huurnink, B., de Rijke, M.: Linking archives using document enrichment and term selection. In: Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries. pp. 360–371. TPDL'11, Springer-Verlag, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2042536.2042584
5. Gottipati, S., Jiang, J.: Linking entities to a knowledge base with query expansion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 804–813. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145523
6. Juric, D., Hollink, L., Houben, G.: Bringing parliamentary debates to the semantic web. In: Proceedings of the workshop on Detection, Representation and Exploitation of Events in the Semantic Web (DeRIVE 2012) (12 November 2012 2012 (to appear))
7. Juric, D., Hollink, L., Houben, G.J.: Discovering links between political debates and media. In: Daniel, F., Dolog, P., Li, Q. (eds.) ICWE. Lecture Notes in Computer Science, vol. 7977, pp. 367–375. Springer (2013), http://dblp.uni-trier.de/db/conf/icwe/icwe2013.html#JuricHH13
8. Kleppe, M., Hollink, L., Kemman, M., Juric, D., Beunders, H., Blom, J., Oomen, J., Houben, G.: Polimedia: Analysing media coverage of political debates by automatically generated links to radio and newspaper items. In: LinkedUp Veni Competition

2013, Proceedings of the LinkedUp Veni Competition on Linked and Open Data for Education. CEUR Workshop Proceedings, vol. 1124, pp. 1–6. CEUR Workshop Proceedings (2014)

9. Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., Chang, Y.: Learning to model relatedness for news recommendation. In: Proceedings of the 20th International Conference on World Wide Web. pp. 57–66. WWW '11, ACM, New York, NY, USA (2011), `http://doi.acm.org/10.1145/1963405.1963417`
10. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK (2008), `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`
11. Meroño-Peñuela, A., Ashkpour, A., v. Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., v. Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web Journal pp. 588–1795 (2014), `http://www.cs.vu.nl/~frankh/postscript/SWJ2014.pdf`
12. Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R.: Linked humanities data: The next frontier? a case-study in historical census data. In: Proc. of the 2nd Int. Workshop on Linked Science 2012. vol. 951 (2012)
13. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg (2013), `http://dx.doi.org/10.1007/978-3-642-28569-1_5`
14. Stasiu, R., Heuser, C., da Silva, R.: Estimating recall and precision for vague queries in databases. In: Pastor, O., Falcão e Cunha, J.a. (eds.) Advanced Information Systems Engineering, Lecture Notes in Computer Science, vol. 3520, pp. 187–200. Springer Berlin Heidelberg (2005), `http://dx.doi.org/10.1007/11431855_14`
15. Téllez-Valero, A., Montes-y Gómez, M., Villaseñor Pineda, L.: A Machine Learning Approach to Information Extraction. pp. 539–547 (2005), `http://www.springerlink.com/content/al7dpnd4a52kg4g9`
16. Yu, B.: An evaluation of text classification methods for literary study. LLC 23(3), 327–343 (2008), `http://dblp.uni-trier.de/db/journals/lalc/lalc23.html#Yu08`