Reverse-Transliteration of Hebrew script for Entity Disambiguation

Aaron Christianson
a.christianson@ub.uni-frankfurt.de
University Library Johann Christian Senckenberg
Frankfurt am Main, Germany

Maral Dadvar
dadvar@hdm-stuttgart.de
Web-based Information Systems and Services (WISS) Stuttgart Media University
Stuttgart, Germany

Kai Eckert
eckert@hdm-stuttgart.de
Web-based Information Systems and Services (WISS) Stuttgart Media University
Stuttgart, Germany

ABSTRACT

JudaicaLink is a novel domain-specific knowledge base for Jewish culture, history, and studies. JudaicaLink is built by extracting structured, multilingual knowledge from different sources and it is mainly used for contextualization and entity linking. One of the main challenges in the process of aggregating Jewish digital resources is the use of the Hebrew script. The proof of materials in German central cataloging systems is based on the conversion of the original script of the publication into the Latin script, known as Romanization. Many of our datasets, especially those from library catalogs, contain Hebrew authors' names and titles which are only in Latin script without their Hebrew script. Therefore, it is not possible to identify them in and link them to other corresponding Hebrew resources. To overcome this problem, we designed a reverse-transliteration model which reconstructs the Hebrew script from the Romanization and consequently makes the entities more accessible.

CCS CONCEPTS

• **Information systems** → **Extraction, transformation and loading**; **Digital libraries and archives**; **Resource Description Framework (RDF)**.

KEYWORDS

Linked Open Data, Retro-Conversion, Transcription, Writing Systems, Digital Library

iiWAS2019, December 2–4, 2019, Munich, Germany
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7179-7/19…$15.00
https://doi.org/10.1145/3366030.3366099

ACM Reference Format:
Aaron Christianson, Maral Dadvar, and Kai Eckert. 2019. Reverse-Transliteration of Hebrew script for Entity Disambiguation . In *The 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019), December 2–4, 2019, Munich, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3366030.3366099

1 INTRODUCTION

A knowledge base is a collection of knowledge about a variety of entities and it contains facts explaining those entities [10]. Besides being used for applications such as question answering [6], and visualization [7], knowledge bases also play an important role in opening up new possibilities for tasks such as Entity Linking; i.e. linking named entity mentions appearing in a text with their corresponding entities in a knowledge base [11]. General-purpose knowledge bases, such as DBpedia[1], are huge resources and using them for domain-specific tasks, when only a small subset of the data is actually relevant, is hard and not very efficient. Therefore, many domain-specific knowledge bases are being developed by gathering resources and information about a specific topic, such as the domain-specific knowledge base for occupations and job activities [5] or JudaicaLink [4], which is a domain-specific knowledge base for Jewish culture and studies, which forms the basis for the work presented in this paper. JudaicaLink is built by extracting structured, multilingual knowledge from different sources, e.g. various online encyclopediae, gazeteers or handbooks. For each source, a separate dataset is created to track data provenance. A main task therefore is the interlinking and identity resolution of entities across these datasets.

One challenge we face in this process is that the datasets not only use different languages, but also different scripts, such as Cyrillic and Hebrew. Often such barriers can be overcome by common cross-lingual matching techniques as exploitation of the entity context or multi-lingual sources (particularly DBpedia) as intermediate link target. There is, however, a substantial amount of library metadata with

[1] http://wiki.dbpedia.org

only little to none context and Hebrew authors' names and titles which are only in Romanized form without the original Hebrew script. Therefore, it is not possible to identify them in and link them to corresponding Hebrew resources easily.

To overcome this problem, we designed a reverse transliteration model which reconstructs the Hebrew script from its Latin script. Having The Hebrew script enables us to find the right match in the Hebrew resources, such as data from the National Library of Israel[2], and link the entities to their corresponding matches. In doing this, we have to deal with name similarities and also name duplicates.

In the remainder of this paper we briefly introduce JudaicaLink, its data sources and infrastructure. Then we will explain the reverse transliteration method that we have developed to deal with Hebrew names in Latin script and the disambiguation procedure for linking the entities to their correct corresponding entries in other resources.

2 JUDAICALINK

JudaicaLink[3] is a domain-specific knowledge base for Jewish culture and studies. One of the main applications of JudaicaLink is to act as a central reference point for the contextualization of metadata of library collections, i.e., entity resolution within and linking of metadata to improve resource access and to provide richer context to the user.

Data sources and Infrastructure

Reference works such as encyclopediae and glossaries function as guides to specific scholarly domains and are therefore highly relevant data sources for JudaicaLink, for example the YIVO Encyclopedia of Jews in Eastern Europe[4]. By using customized web scrapers, we extract structured data from the article pages and turn every article into a linked and described entity. Furthermore, subsets of other data sources are added such as the Integrated Authority File (GND) of the German National Library[5], the DBpedia or the Virtual Internet Authority File (VIAF) [6]. These sources not only provide additional data, they are also widely acknowledged as authority data and therefore form linking targets for library data from all over the world.

JudaicaLink is constantly developed and extended according to the requirements at hand in our digital library contextualization projects, such as the Specialized Subject Service Jewish Studies[7]. Besides data on persons, it contains mainly data on geographic places and events.

All the code for extraction and data generation is available open source via GitHub[8]. The extracted information is available as Linked Open Data (LOD) using the JudaicaLink ontology[9] and Pubby (DM2E version) as Web Frontend [1]. Dumps are available in Turtle (Terse RDF Triple Language, TTL)[10] from our website[11]. Internally, the datasets are organized as different named graphs. Links between the datasets are mostly generated automatically and form datasets (linksets) on their own. The complete knowledge graph can be accessed as union graph contaning all datasets via a public SPARQL endpoint[12].

The Challenge of the Hebrew Script

There are several challenges that arise while constructing the datasets: unstructured data sources like online encyclopediae need to be made available as structured data with stable URIs, relevant subsets of general-purpose knowledge bases have to be identified to fill the gaps between the specialized resources and to provide further context, and last, to further enrich our datasets, all data sources will be integrated and interlinked, and the identified entities (persons) will be linked to other resources which have further information about these entities.

One of the main challenges in the process of linking Jewish digital resources is the use of the Hebrew script. Today, with the ubiquity of Unicode and the support for right-to-left languages in modern software, dealing with the script itself is trivial. There is a savor of irony that the greatest challenges with Hebrew script today arise from measures taken by previous generations of curators to avoid using it because they could not handle metadata in scripts not known to them prior to the use of digital catalogs. As computers took over, the difficulty of storing and presenting resources containing multiple scripts remained until Unicode came to prominence, so many cataloging systems have standardized on the conversion of the original script into Latin script, a process known as Romanization. For researchers, the absence of Hebrew script in the metadata is a detriment. Researchers may not be familiar with Romanization systems used by a resource. Additionally, creating queries and sifting through results in Romanization is more difficult than using native script. Marquardt has written about these challenges and the status of Romanized Hebrew metadata in various institutions at great length [8]. Weinberg has also done an outstanding job detailing the approach to Hebrew metadata specifically in

[2] http://web.nli.org.il/sites/nlis

[3] http://www.judaicalink.org

[4] http://yivoencyclopedia.org

[5] http://dnb.de

[6] Virtual International Authority File, Available: http:/viaf.org/viaf/data (accessed on 13. September 2018)

[7] Fachinformationsdienst Juedische Studien, http://jewishstudies.de

[8] https://github.com/wisslab/judaicalink-loader

[9] https://github.com/wisslab/judaicalink-ontology/blob/master/judaicalink-ontology.md

[10] https://w3.org/TR/turtle/

[11] http://www.judaicalink.org/datasets/

[12] http://www.judaicalink.org/search/

the USA [12]. Today's curators have recognized these problems, and more recent records for Hebrew entities contain both Hebrew and Romanized forms. However, a great bulk of older Hebrew metadata exist only in Romanized form and are difficult to use or integrate them for other purposes. Therefore, we have developed a system to reconstruct Hebrew from Romanized text and, by integrating the Hebrew script into our datasets which only had the Romanized form, made the entities in these datasets more accessible and therefore possible to be linked to other resources.

3 RECONSTRUCTION OF THE HEBREW FROM THE ROMANIZATION

Relatively little has been written about the conversion of Romanized Hebrew back to the original script. Typing common Hebrew Romanization into Hebrew Google translate will produce suggestions in Hebrew letters, though we have been unable to find any information about their methodology. There are also various attempts which are mostly presented as interactive demos, but again there is no information about their methodology and their performance quality[13]. Ornan and Leket-Mor have written about idealized Hebrew Romanization schemes that would be both reversible and dialect agnostic [9]. Unfortunately, such "ideal" systems do not exist in real libraries. Bar and Dershowitz have done some excellent work on converting Judeo-Arabic (Arabic written in Hebrew letters) to standard Arabic and, while not strictly the same problem, it also deals with script conversion. Their use of statistics for determining correct words has been instructive for our approach [2]. To our knowledge, there are no publications aside from our own [3] that deal with the conversion of Romanization into the original script from real-world metadata. Reconstruction is non-trivial for two reasons. First, none of the several systems of Romanization in our local dataset can be deterministically converted to the source script. In addition, Hebrew orthography is also a source of ambiguity. Many words (and names not least) have multiple traditional spellings. Our approach can be broken into two steps:

(1) Deterministically generate every possible Hebrew form.
(2) Filter out implausible results and rank the remaining forms, singling out the most probable result.

Form Generation

First, the Romanized word is split into tokens using our Python library, *Deromanize*. Additional documentation about how the Romanization standard is defined is available at the project repository[14]. Each token corresponds to one or more possible Hebrew conversion forms. The conversions are created from predefined lists where more likely conversions are given a higher rank.[15]

sh	*a*	*l*	*o*	*m*
0 ש	0 -	0 ל	0 ו	0 ם
	10 א		3 -	
			10 א	

Figure 1: tokens and possible conversions

Next, the Cartesian product over all conversion forms is used to create all hypothetical Hebrew forms for the given Romanized form. A normalized rank for each Hebrew form is calculated based on the combined rank of the single tokens. Generating these products could theoretically be a bottleneck, but it has not been in practice.

0.64	0.16	0.06	0.06	0.05	0.03
שלום	שלם	שלאם	שאלום	שאלם	שאלאם

Figure 2: tokens combined

Filtering

After all the forms have been generated by Deromanize, we move on with filtering and further prioritizing these generated forms. Our initial work in this area was focused primarily on the conversion of book titles [3] and now we have expanded our work to conversion of authors names. In the initial study, several filtering steps were tried for selecting the best results. The simplest approach was to spell-check to find the forms that are real words. As our earlier study details, spell checking by itself is not reliable when multiple real words are produced. In the example in Figure 2, the top result is the correct Hebrew form, but the second suggestion is also a real word, *shalem.* In this case, the top suggestion would still be correct because we factor in the predefined ranks, but this is unreliable. Additionally, foreign words, traditional spellings and personal names are not in a spelling dictionary. We therefore created a specialized spelling dictionary of Hebrew personal names based on name elements from the NLI authority data, as well as Hebrew data from VIAF. This dictionary contains each name element and the number of times it occurs in the datasets. However, personal names are still problematic because many orthographic variations can exist

[13] https://stevemorse.org/hebrew/eng2heb.html, http://www.alittlehebrew.com/transliterate/, http://transliterate.com

[14] https://github.com/FID-Judaica/deromanize

[15] Note that this is not a probabilistic approach; Ranks are determined in predefined lists which have been created by domain experts.

for names which sound the same and therefore often also share a Romanized form. In the case of Jewish names, they may have Hebrew spellings as well as additional spellings in Hebrew characters for the local Jewish dialects, such as Yiddish, Ladino or Judeo-Arabic. The VIAF dataset was further used to create a lookup table which contains forms of the full name in other languages correlated to their Hebrew forms. This is useful because our local authority data often also has forms of the name in other languages. For example, if we have the name "Mosheh ben Asher", *Mosheh*, *ben* and *Asher* are all considered separate elements. We use the name dictionary to verify generated forms for each element. A Cartesian product is made of the top four forms for each element, and each combination is tested against the results from the full-name lookup, and those that match become the top candidates. We then match the top three candidates directly against names found in the NLI dataset.

4 RESULTS

For the evaluation of our reverse transliteration, we used one of our multi-lingual datasets[16] which contains 5,170 person names both in Romanization and Hebrew script. Using our approach, we produced the top-ranked Hebrew form for each Romanized form and compared the result to the Hebrew version in our data. With 3,878 matches, we reached an accuracy of 75%.

Besides the names with Hebrew form, our dataset also contains 9,474 person's names where only the Romanized form is available. To get a better understanding of the applicability for entity disambiguation in our project, we used the top three candidate forms to search for matches for these names in NLI authority data[17]. This way, at least one matching entity could be obtained for 7,333 names (77.4%); 2,534 of these names had more than one match in the NLI data. For disambiguation, we considered additional personal information such as date of birth and death, where available. This way, we were able to select a single matching entity for 1085 records (42.8% of the duplicate names).

5 CONCLUSION

In this paper, we presented a novel approach for reverse transliteration of Romanized Hebrew for the use case of entity matching. With 75% correct reversions and over 77% of matching candidates, we certainly have a good basis for entity matching. Nevertheless, we want to focus on the missing 25% first: based on the results of the matching, we will further refine the predefined lists and therefore the prioritization of our guesses. In contrast to general Hebrew texts, that we used as starting point [3], the case of personal names presents special challenges and follows own rules. Regarding the disambiguation and the actual entity matching, we will investigate how existing information can be best used and where we can obtain further context by bridging the script barrier using our approach. Besides the existing open source release, we also intend to make all of our tools available for external use via a public API, so that the approach can easily be tested on other data.

REFERENCES

[1] Konstantin Baierer, Evelyn Dröge, Kai Eckert, Doron Goldfarb, Julia Iwanowa, Christian Morbidoni, and Dominique Ritze. 2017. DM2E: A linked data source of digitised manuscripts for the digital humanities. *Semantic Web* 8, 5 (2017), 733–745.

[2] Kfir Bar, Nachum Dershowitz, Lior Wolf, Yackov Lubarsky, and Yaacov Choueka. 2016. Processing Judeo-Arabic Texts. (2016).

[3] Aaron Christianson, Rachel Heuberger, and Thomas Risse. 2018. Back to the Source: Recovering Original (Hebrew) Script from Transcribed Metadata. In *Digital Libraries for Open Knowledge*, Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes (Eds.). Springer International Publishing, Cham, 326–329.

[4] Maral Dadvar and Kai Eckert. 2018. JudaicaLink; A Domain-Specific Knowledge Base for Jewish Studies. In *17th Dutch-Belgian Information Retrieval Worksop (DIR2018)*. Leiden, Netherlands. http://arxiv.org/abs/1812.04265

[5] Jonas Bulegon Gassen, Stefano Faralli, Simone Paolo Ponzetto, and Jan Mendling. 2016. Who-Does-What: A Knowledge Base of People's Occupations and Job Activities.. In *International Semantic Web Conference (Posters & Demos)*.

[6] Yoji Kiyota, Sadao Kurohashi, and Fuyuko Kido. 2002. Dialog navigator: A question answering system based on large text knowledge base. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1–7.

[7] Peter Kraker, Christopher Kittel, and Asura Enkhbayar. 2016. Open Knowledge Maps: Creating a Visual Interface to the WorldâĂŹs Scientific Knowledge Based on Natural Language Processing. *027.7 Zeitschrift für Bibliothekskultur/Journal for Library Culture* 4, 2 (2016), 98–103.

[8] Susanne Marquardt. 2005. Transliteration und Retrieval. Zur Problematik des Auffindens hebräischsprachiger Medien in Online-Katalogen. *Berliner Handreichungen zur Bibliothekswissenschaft* 157 (2005).

[9] Uzzi Ornan and Rachel Leket-Mor. 2016. Phonemic Conversion as the Ideal Romanization Scheme for Hebrew: Implications for Hebrew Cataloging. *Judaica Librarianship* 19 (2016), 43–72.

[10] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International Semantic Web Conference*. Springer, 177–185.

[11] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2015), 443–460.

[12] Bella Hass Weinberg. 1990. Automation and the American Judaica Library during the First Quarter Century of the Association of Jewish Libraries, 1965-1990. *Judaica Librarianship* 5 (1990), 167–176.

[16] http://www.judaicalink.org/datasets/ubffm-persons

[17] 182,447 persons, http://www.judaicalink.org/datasets/nli-persons