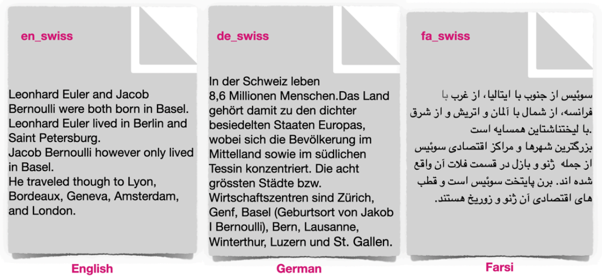
From unstructured texts to RDF-star-based open research data queryable by references

Alassi, Sepideh
sepideh.alassi@unibas.ch
University of Basel, Switzerland

Automated extraction of information from text and its transformation into machine readable data enables efficient query of information. Strategies have been presented for extraction and linking of named entities from texts with knowledge graph individuals, and their association with grammatical units that lead to producing more coherent facts (Martinez-Rodriguez et al 2018, p.339). These strategies rely on approaches using Information Extraction to populate the Semantic Web and/or using Semantic Web resources to improve Information Extraction (Martinez-Rodriguez, et.al 2016). Projects like LODifier take the unilateral approach to populate a knowledge graph based on automatic extraction of named entities and their relations from unstructured English texts (Augenstein et all 2012). Our project takes the bilateral approach to semantically link text documents in different languages that have references in common. This would make the texts queryable by references irrespective of the language of the text. Moreover, this project attempts to extract the relations between named entities to augment the knowledge graph. Metadata such as the source of the information can be expressed as a statement about a statement using RDF-star technology; adding the metadata to the edges of the graph. This is particularly useful for citation as well as metadata-based query of data without dealing with deficits of standard RDF such as reification (Alassi, Rosenthaler 2022). This project focuses on extracting and linking three entries from unstructured texts in different languages: locations, persons, and their relations. The project workflow is described using the texts in English, German, and Persian in Figure 1.


en_swiss

Leonhard Euler and Jacob Bernoulli were both born in Basel.
Leonhard Euler lived in Berlin and Saint Petersburg.
Jacob Bernoulli however only lived in Basel.
He traveled though to Lyon, Bordeaux, Geneva, Amsterdam, and London.

English

de_swiss

In der Schweiz leben 8,6 Millionen Menschen.Das Land gehört damit zu den dichter besiedelten Staaten Europas, wobei sich die Bevölkerung im Mittelland sowie im südlichen Tessin konzentriert. Die acht grössten Städte bzw. Wirtschaftszentren sind Zürich, Genf, Basel (Geburtsort von Jakob I Bernoulli), Bern, Lausanne, Winterthur, Luzern und St. Gallen.

German

fa_swiss

سوئیس از جنوب با ایتالیا، از غرب با فرانسه، از شمال با آلمان و اتریش و از شرق با لیختناشتاین همسایه است. بزرگترین شهرها و مراکز اقتصادی سوئیس از جمله ژنو و بازل در قسمت فلات آن واقع شده اند. برن پایتخت سوئیس است و قطب های اقتصادی آن ژنو و زوریخ هستند.

Farsi

Figure 1. Test texts.

An RDF-star based ontology is defined for this project using the existing ontologies such as **foaf**[1] and **dbo**[2] by making subclasses and sub-properties, see Figure 2.

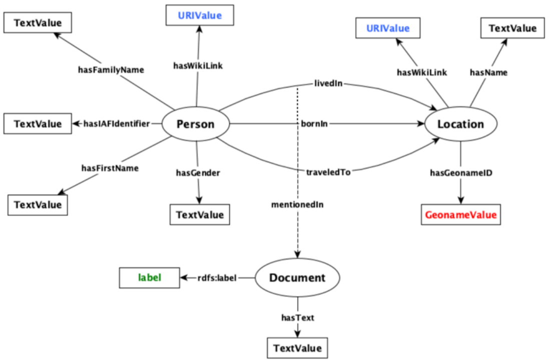


Figure 2. Ontology.

Figure 3 illustrates the first stage of the workflow where the references to locations and persons are extracted from the input texts using named entity recognition (NER) processes and pretrained language models. For each recognized entity, information required by the ontology are retrieved from Wikidata. GND numbers for persons and Geoname IDs for locations are used to unify the entities given in different languages. The results of the NER and information retrieval from Wikidata are verified and missing identifiers are added. The knowledge graph is then populated with extracted information creating resources for location and person entities. Next, references to entities are replaced in texts with standoff links to the corresponding resources using their IRIs and documents are stored with enriched textual bodies.

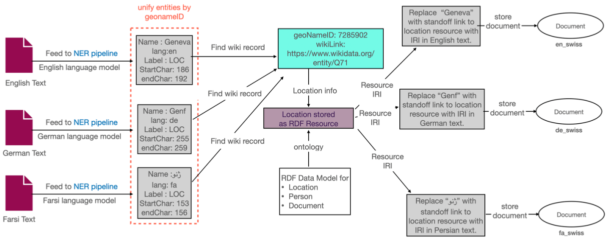
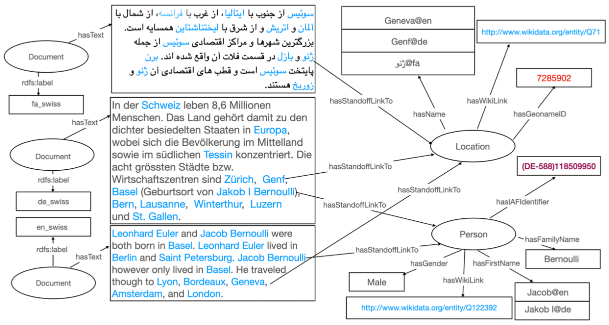


Figure 3. Workflow of stage 1.

In this way, documents with references in common are linked, see Figure 4. Now we can query for all documents mentioning a specific person or a location. For example, a query for texts with reference to a location with name “Geneva” would return all documents in which this location is referenced in any language “Ge-neva”, “Genf”, “ژنو”, etc.

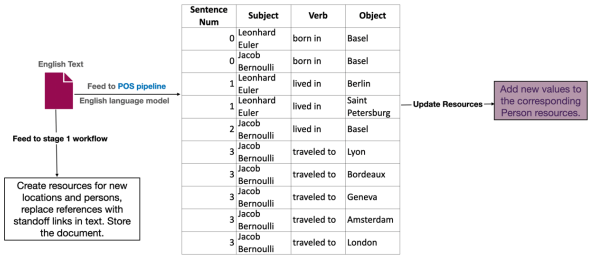

سوئیس از جنوب با ایتالیا، از غرب با فرانسه، از شمال با آلمان و اتریش و از شرق با لیختناشتاین همسایه است. بزرگترین شهرها و مراکز اقتصادی سوئیس از جمله ژنو و بازل در قسمت فلات آن واقع شده اند. برن پایتخت سوئیس است و قطب های اقتصادی آن ژنو و زوریخ هستند.

In der Schweiz leben 8,6 Millionen Menschen. Das Land gehört damit zu den dichter besiedelten Staaten in Europa, wobei sich die Bevölkerung im Mittelland sowie im südlichen Tessin konzentriert. Die acht grössten Städte bzw. Wirtschaftszentren sind Zürich, Genf, Basel (Geburtsort von Jakob I Bernoulli), Bern, Lausanne, Winterthur, Luzern und St. Gallen.

Leonhard Euler and Jacob Bernoulli were both born in Basel. Leonhard Euler lived in Berlin and Saint Petersburg. Jacob Bernoulli however only lived in Basel. He traveled though to Lyon, Bordeaux, Geneva, Amsterdam, and London.

Figure 4. Excerpt of the graph.

In the second stage of the workflow (Figure 5) semantic web resources are utilized to improve extraction of the relations between named entities (locations and persons). Based on the defined ontology, dependency graphs of sentences, and part of speech (POS) tags, the relations between named entities are extracted from the text. The definitions of resource classes and predicates together with their subject and object class constraints are considered as parsing rules. For example, predicate “ **livedIn**” has subject type **Person** and object type **Location**, thus, only relations are considered where subject of the sentence is a person entity, object (of proposition) is a location entity, and the lemmatized form of the verb is “live” with the proposition “in”. The gender information of persons are used for pronoun resolution. The resources representing the named entities of a sentence are put in a LIFO stack to be used for backwards resolution of personal pronouns in the succeeding sentence.


English Text
Feed to POS pipeline
English language model
Feed to stage 1 workflow
Create resources for new locations and persons, replace references with standoff links in text. Store the document.
Update Resources
Add new values to the corresponding Person resources.

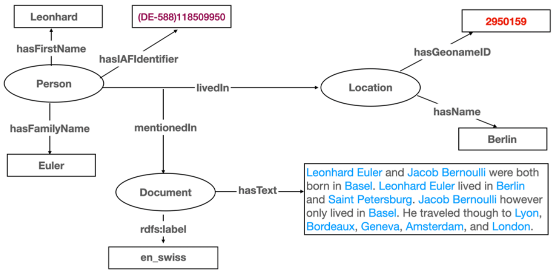
Sentence Num	Subject	Verb	Object
0	Leonhard Euler	born in	Basel
0	Jacob Bernoulli	born in	Basel
1	Leonhard Euler	lived in	Berlin
1	Leonhard Euler	lived in	Saint Petersburg
2	Jacob Bernoulli	lived in	Basel
3	Jacob Bernoulli	traveled to	Lyon
3	Jacob Bernoulli	traveled to	Bordeaux
3	Jacob Bernoulli	traveled to	Geneva
3	Jacob Bernoulli	traveled to	Amsterdam
3	Jacob Bernoulli	traveled to	London

Figure 5. Workflow of stage 2.

The source document is added to the edges of the graph through **mentionedIn** predicate (Figure 6). RDF-star triple below represents the source of the triple :euler :livedIn :berlin.

<<:euler :livedIn :berlin >> :mentionedIn :en_swiss .

Through SPARQL-star, one can then query for documents containing a certain relationship between entities.



Leonhard Euler and Jacob Bernoulli were both born in Basel. Leonhard Euler lived in Berlin and Saint Petersburg. Jacob Bernoulli however only lived in Basel. He traveled though to Lyon, Bordeaux, Geneva, Amsterdam, and London.

Figure 6. Excerpt of the RDF-star Graph.

The first stage of the workflow is configurable to use either *flair*[3] or *spaCy*[4] for NER. However, spaCy is used for dependency parsing because it utilizes Stanford-type dependency terminology (de Marneffe, Manning 2016). The second stage of the workflow currently supports only English texts and will be extended to German texts.

Notes

[1] http://xmlns.com/foaf/0.1/.
[2] https://www.dbpedia.org/resources/ontology/.
[3] https://github.com/flairNLP/flair
[4] https://spacy.io/

Bibliography

Augenstein, I., Padó, S., Rudolph, S.: LODifier: Generating linked data from unstructured text. In: The Semantic Web: Research and Applications. Volume 7295 of LNCS. Springer Berlin (2012) 210–224

Alassi, Sepideh, Rosenthaler Lukas, “RDF-star-based Digital Edition of Travel Journals.”, DH2022 Tokyo, 2022.

Martinez-Rodriguez, Jose L., Ivan López-Arévalo, and Ana B. Rios-Alvarado. "Openie-based approach for knowledge graph construction from text." Expert Systems with Applications 113 (2018): 339-355.

Martinez-Rodriguez, Jose L., Aidan Hogan, and Ivan López-Arévalo. "Information Extraction meets semantic web: A survey.”

De Marneffe Marie-Catherine, Christopher D. Manning, “Stanford typed dependencies manual”, 2016. Available at https://downloads.cs.stanford.edu/nlp/software/dependencies_manual.pdf