An interactive AI-based approach to semantic annotations for the SpokenWeb archive.

About the author

Francisco Berrizbeitia Eng, M.Sc is a developer at Concordia Library and the lead developer of Swallow. His interests lie in linked open data, text mining, and natural language understanding, currently collaborating with researchers from multiple institutions.

Introduction

Adding semantic annotations to archival metadata allows to generate an alternative representation of the dataset in the form of a graph. This can be useful for multiple reasons: discovery of new relationships between objects, improves findability and allows for more sophisticated queries using the sparql query language. In this poster we will explain the rationale used to develop a web-based tool to help users deal with this task using a semi-automatic approach that ensures high quality annotations while leveraging natural languages understanding techniques to speed up the process.

Research question

Can a process, based on a general NER pretrained model effectively tag descriptions of literary audio for cataloguing purposes?

Experimental design

To test the hypothesis we first designed and implemented an automated process and then compared the results to a manually-tagged dataset (the gold standard).

The manually-tagged data set we used was that of the Sir George Williams Poetry Series, consisting of 54 unique entries in Swallow documenting twice as many recorded events, with entries sometimes having as many as 30 or more Wikidata q-codes per unique item content field.

Automatic tagging process

1. Preprocessing
    1.1. Extract the manually tagged entities from the text.
    1.2. Generate a list with the entities (gold standard) and a clean version of the text to be passed to Dbpedia Spotlight
    1.3. Save the results in a new file.
2. NER
    2.1. Process the clean text using the Dbpedia Spotlight API. This results in a list of dbpedia.org links.
    2.2. Access the dbpedia.org link and look for a Wikidata.org equivalent defined via the “sameAs” predicate. If one is found, add the Wikidata URL to a list.
    2.3. Save the results in a new file.

3. Analysis
    3.1. Compare both lists and calculate for each record: a) number of entities on the gold standard, b) true positives, c) false positives, d) precision, and e) recall.
    3.2. Save the results on a new file.

Results

Default Parameters
True Positives: 665
False Positive: 990
Recall: 0.48
Precision: 0.40

Addition of filters
True Positives: 501
False Positive: 354
Recall: 0.36
Precision: 0.59

Looking closely at the false positives we found out that many entities weren't wrong but considered irrelevant by the cataloguer. If we accept this entities the precision increases to 80%

Analysis

We consider these results encouraging enough, not to fully automate the process, but to make useful suggestions that could make the annotation process faster for the cataloguers.

With this in mind, we then proceeded to develop a web application that could be integrated with Swallow or be used independently. The application uses a python back end that takes care on the interactions with dbpedia-spotlight and Wikidata.org and exposes the different methods as web services using Flask. The front-end is an easy to use, JavaScript based user interface depeticted below.

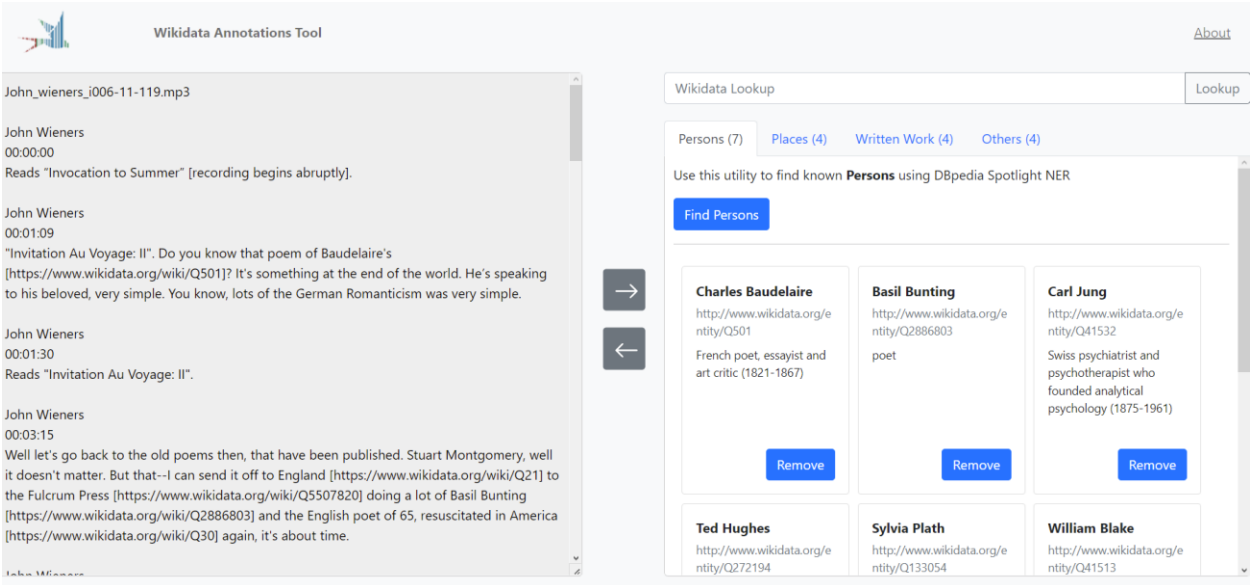

Figure 1. UX design of the annotation tool

We hope that tools like the one we are proposing will encourage catalogue administrators to include semantic annotations in the records and connect more collections to the linked data cloud.

To read a more detailed explanation of the experiment please visit: https://spokenweb.ca/a-proposal-for-semantic-annotations-an-ai-assisted-approach/