HHS Public Access

Author manuscript

Proc IEEE Int Conf Semant Comput. Author manuscript; available in PMC 2022 March 04.

Published in final edited form as:

Proc IEEE Int Conf Semant Comput. 2021 January ; 2021: 88–89. doi:10.1109/icsc50631.2021.00022.

Extracting Semantics from Census-based Reference Data

Daniel R. Harris,

Center for Clinical and Translational Sciences, Institute for Pharmaceutical Outcomes and Policy, University of Kentucky, Lexington, KY USA

Nima Seyedtalebi

Department of Computer Science, College of Engineering, University of Kentucky, Lexington, KY USA

Abstract

We present preliminary findings in extracting semantics from reference data generated by the United States Census Bureau. US Census reference data is based upon surveys designed to collect demographics and other socioeconomic factors by geographical regions. These data sets contain thousands of variables; this complexity makes the reference data difficult to learn, query, and integrate into analyses. Researchers often avoid working directly with US Census reference data and instead work with census-derived extracts capturing a much smaller subset of records. We propose to use natural language processing to extract the semantics of census-based reference data and to map census variables to known ontologies. This semantic processing reduces the large volume of variables into more manageable sets of conceptual variables that can be organized by meaning and semantic type.

Index Terms—

natural language processing; semantic technology

I. Introduction

In the United States, large scale census surveys such as the Decennial Census [1] and the American Community Study (ACS) [2] collect large amounts of demographic and socioeconomic data that ultimately are aggregated at different levels of census-designated geographical boundaries. The ACS data is divided into several key parts: detailed tables (over 20,000 variables), subject tables (over 16,000 variables), data profiles (over 1,000 variables), and comparison profiles (over 1,000 variables) [2]. Current perception from researchers using the ACS maintains that there exists a difficult learning curve due to the complexity and vastness of the data [3]. The ACS is a *wide* data set having an incredible number of columns which poses problems for users attempting to use the data; finding the correct column or even knowing if it exists can be problematic for those unfamiliar with the ACS [3], [4]. The large number of columns also poses technical issues due to commonly

daniel.harris@uky.edu .

found limits on the number of columns allowed in a single table within relational database systems [5].

We apply natural language processing (NLP) techniques for concept extraction to explore the implicit semantics of census data. NLP on short texts has shown that concept extraction can be effective for prescription instructions [6], prescription indications [7], [8], and radiology figure captions [9]. We extract and integrate ad-hoc demographic and socioeconomic factors from the census to support ad-hoc research requests and data dashboards for the HEALing Communities Study and evaluating the Communities That Heal (CTH) intervention on reducing opioid overdose deaths in 67 disproportionately affected communities [10]. To our knowledge, this is the first effort to extract and utilize the semantics of ACS data, despite it being a critical data set for socioeconomic and population health research.

II. Methods and Discussion

In order to extract semantic information from census data, we use MetaMap [11], a concept extraction tool based on both NLP and computational-linguistic techniques. MetaMap was designed to process biomedical libraries and clinical text, and while census data is not primarily biomedical, many of the same concepts needed to describe patient population demographics also apply to a much more general citizen population. Using the documentation for the Census API [12], we downloaded meta-data for two collections within the ACS, subject tables (over 16,000 variables) and data profiles (over 1,000 variables) [2]. We parsed these meta-data files to extract table names, official census variable names, and a flattened list of unique components from the variable's label. Census variable labels are a delimited list separated by *!!* symbols. For example, in *SELECTED ECONOMIC CHARACTERISTICS*, the variable *DP03_0130E* has the label *Estimate!! PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL!!Under 18 years!!Related children of the householder under 18 years*. We split long labels into an item set:

1. Estimate
2. PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL
3. Under 18 years
4. Related children of the householder under 18 years

We assign an ID number to each of the unique items; this unique list of items is then processed with MetaMap using the term processing feature, restricted to SNOMED CT concept mappings, and with word-sense disambiguation (WSD) enabled. WSD will select the best mapping amongst candidates, resulting in a "best fit" mapping between items in our census variable labels and concepts in SNOMED CT. Census variables have a 1-many relationship with the item list, so effectively census variables also have a 1-many relationship with concepts.

Table I shows the DP03 0130E variable discussed earlier and the resulting semantic mappings from our process; this one variable contains four items which map to twelve concepts. Our output also contains the relevance score and offset of the original text that triggered the concept. The score assesses the concept's relevance to the text. In our example, *poverty* is scored significantly higher than the other concepts, followed by *child*, *income*, and *estimated*, respectively.

Basic counts for data profiles and subject tables are summarized in Table II, where items are the pieces parsed out of the variable label and converted into a concept within the UMLS. The counts presented represent unique counts; the ACS Subject Tables had 16,558 unique variables containing 1,284 unique items which map to 423 unique concepts across 58 unique semantic types.

The benefit of performing concept extraction on the items parsed from the original census variables is that semantically similar items can be retrieved together, despite heavy lexical variation. For example, in our results, items *occupied housing units* and *housing occupancy* share the same concept unique identifier (CUI) in SNOMED CT. A search matching one of these housing-related items can easily be expanded to reach the other because they are semantically linked together with the same CUI. An additional example is that *unemployed* shares a concept with *unemployment rate*, despite being phrased significantly different.

III. Conclusion and Future Work

We demonstrated that natural language processing tools can be used to extract semantics from the US Census data. We are evaluating the effectiveness of our concept extraction and its ability to assist in ad-hoc data requests and dashboards for the HEALing Communities Study [10]. In particular, we plan to demonstrate the utility of semantic linkages amongst concepts generated from our analyses. We wish to provide open-source code for semantic processing and querying of reference data so that others may benefit from this effort.

Acknowledgment

The project described was supported by the National Institutes of Health through the NIH HEAL Initiative under award number UM1DA049406 and the National Center for Advancing Translational Sciences through grant number UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1]. 2020 census. Accessed Oct. 1, 2020. [Online]. Available: https://2020census.gov/en.html

[2]. American community survey (acs). Accessed Oct. 1, 2020. [Online]. Available: https://www.census.gov/programs-surveys/acs

[3]. Donnelly FP, "The american community survey: practical considerations for researchers," Reference services review, vol. 41, no. 2, pp. 280–297, 2013.

[4]. Hayslett M and Kellam L, "The american community survey: Benefits and challenges," IASSIST Quarterly, vol. 33, no. 4, pp. 31–31, 2010.

[5]. Bhagat V and Gopal A, "Comparative study of row and column oriented database," in 2012 Fifth International Conference on Emerging Trends in Engineering and Technology. IEEE, 2012, pp. 196–201.

[6]. Harris DR, Henderson DW, and Corbeau A, “sig2db: a workflow for processing natural language from prescription instructions for clinical data warehouses,” AMIA Summits on Translational Science Proceedings, vol. 2020, p. 221, 2020.

[7]. Khare R, Li J, and Lu Z, “Labeledin: cataloging labeled indications for human drugs,” Journal of biomedical informatics, vol. 52, pp. 448–456, 2014. [PubMed: 25220766]

[8]. Khare R, Wei C-H, and Lu Z, “Automatic extraction of drug indications from fda drug labels,” in AMIA Annual Symposium Proceedings, vol. 2014. American Medical Informatics Association, 2014, p. 787. [PubMed: 25954385]

[9]. Kahn CE Jr and Rubin DL, “Automated semantic indexing of figure captions to improve radiology image retrieval,” Journal of the American Medical Informatics Association, vol. 16, no. 3, pp. 380–386, 2009. [PubMed: 19261938]

[10]. Wu E, Villani J, Davis A, Fareed N, Harris DR, Huerta TR, LaRochelle MR, Miller CC, and Oga EA, “Community dashboards to support data-informed decision making in the healing communities study,” Drug and Alcohol Dependence, p. 108331, 2020. [PubMed: 33070058]

[11]. Aronson AR, “Metamap: Mapping text to the umls metathesaurus,” Bethesda, MD: NLM, NIH, DHHS, pp. 1–26, 2006.

[12]. Available APIs. Accessed Oct. 1, 2020. [Online]. Available: https://www.census.gov/data/developers/data-sets.html

TABLE I

Example semantic mappings for census variable DP03_0130E from Data Profiles

Item	Concept Name	Semantic Type
estimate	estimated	Quantitative
percentage of families and people whose income in the past 12 months is below the poverty level	percent (qualifier value)	Quantitative
	family member	Family Group
	income	Quantitative
	in the past	Temporal
	month	Temporal
	poverty	Group Attribute
	levels (qualifier value)	Qualitative
under 18 years	year	Temporal
related children of the householder under 18 years	related personal status	Finding
	child	Age Group
	year	Temporal

TABLE II

Overview of Census Data Profiles (DP) and Subject Tables (ST)

Data Set	Variables	Items	Concepts	Semantic Types
DP	1,375	534	295	47
ST	16,558	1,284	423	58