A Lexico-semantic System for Survey Variable Search

Keywords: information retrieval, survey data search, survey data harmonization, semantic textual similarity, natural language processing

Extended Abstract

Background. In the social sciences, comparing survey data from different sources is important because it allows the study of social phenomena over time and across countries [1, 2]. However, many challenges are associated with searching and comparing across different surveys, such as choosing indicators, the wording of the questions, questionnaire design, and response categories. Through survey data *harmonization* [3], the comparability of survey data can be improved (e.g., by finding similar questions or recoding different data into a unified format). *Ex-post harmonization* deals with finding similar or identical questions across surveys to compare existing data not originally collected for comparative purposes [3]. Existing tools for finding questions, such as GESIS Search[1] or Roper's iPoll,[2] use simple keyword matching algorithms. An obvious limitation is that a user must know exact keywords or be familiar with relevant survey data to use such tools. This prerequisite acts as a barrier of entry for early-stage researchers and hinders access to research data along the FAIR Principles.

Methodology. We present a semantic search engine for survey questions in the form of a web application. The search engine incorporates recent advances in the fields of Natural Language Processing (NLP) and Information Retrieval (IR) by combining lexical search (i.e., keyword search) with semantic search into a hybrid system.

We use BM25 as the lexical model to match keywords and E5 [4], a text embedding model, for semantic matching of phrases and sentences. BM25 is a strong lexical baseline while E5 is a recently-released model that was trained in a contrastive learning setup on 1.3 billion text pairs and is suitable for retrieval and clustering tasks. During retrieval, rather than computing the similarity of the input with each unique question, we use approximate nearest neighbor search using the FAISS library [5], allowing the system to scale to millions of questions potentially. The system takes as input any sequence of words (e.g., a phrase or a sentence) and retrieves lexically and semantically similar results. Multiple variables are grouped if they have the same question. This helps foster ex-post harmonization across surveys, allowing users to identify multiple comparable variables from different surveys quickly. At the same time, the question associated to each group can be used to find comparable variables that are syntactically different but semantically similar to the search input. While the current system only works in English, multilingual models can easily be integrated to provide cross-lingual semantic search [6].

Dataset. We collect variables in English from the GESIS index, filtering variables with no question texts and removing duplicates with respect to all variable metadata, resulting in over 17,000 unique questions and over 83,000 variables from 202 studies. The studies were conducted between 1953 and 2017, most of which are an iteration of the Eurobarometer.

[1] `https://search.gesis.org/`

[2] `https://ropercenter.cornell.edu/ipoll/`

User interface. A screenshot of the user interface is shown in Figure 1. Besides the main text search field, we implement simple filters, such as a the country where the study was conducted and the year. The results are ranked in order of lexical overlap and semantic similarity to the search input based on the underlying search models. Identical variables are grouped together, and variable-level details are shown when a grouping is opened.

Evaluation. We evaluated our system using the variable correspondence list of the ISSP 'Social Inequality I-IV' cumulation, which maps similar variables from four years of the module to a single unified variable. Using the question for a given year, we retrieve the top 10 most relevant unique questions and test the ability of our system to retrieve similar variables. We evaluate the performance of our hybrid system using (Mean) Average Precision (MAP), a standard performance metric used in Information Retrieval, with respect to the top 10 most relevant results. Our system achieves a MAP score of 0.29, indicating that it can retrieve relevant documents close to 30% of the time. In the future, we plan to conduct systematic evaluations, including user-based testing, to validate our findings further. Furthermore, we will incorporate multilingual search to make the system more inclusive to other languages and train custom text embedding models.

Demo walkthrough. The demo will consist of two main parts: 1) an *interactive system demonstration* where we will first guide the audience through the main functionality of the search engine and the underlying searchable dataset and then let researchers input their queries and discuss the results together (crucially including current limitations of the system); 2) an *overview poster* summarizing the main methodological components of our system (e.g., technical details on the lexical and semantic search models), as well as the quality of the search based on a user-based evaluation.

References

[1] Thomas Krämer, Andrea Papenmeier, Zeljko Carevic, Dagmar Kern, and Brigitte Mathiak. Data-seeking behaviour in the social sciences. *Int. J. Digit. Libr.*, 22(2):175–195, 2021.

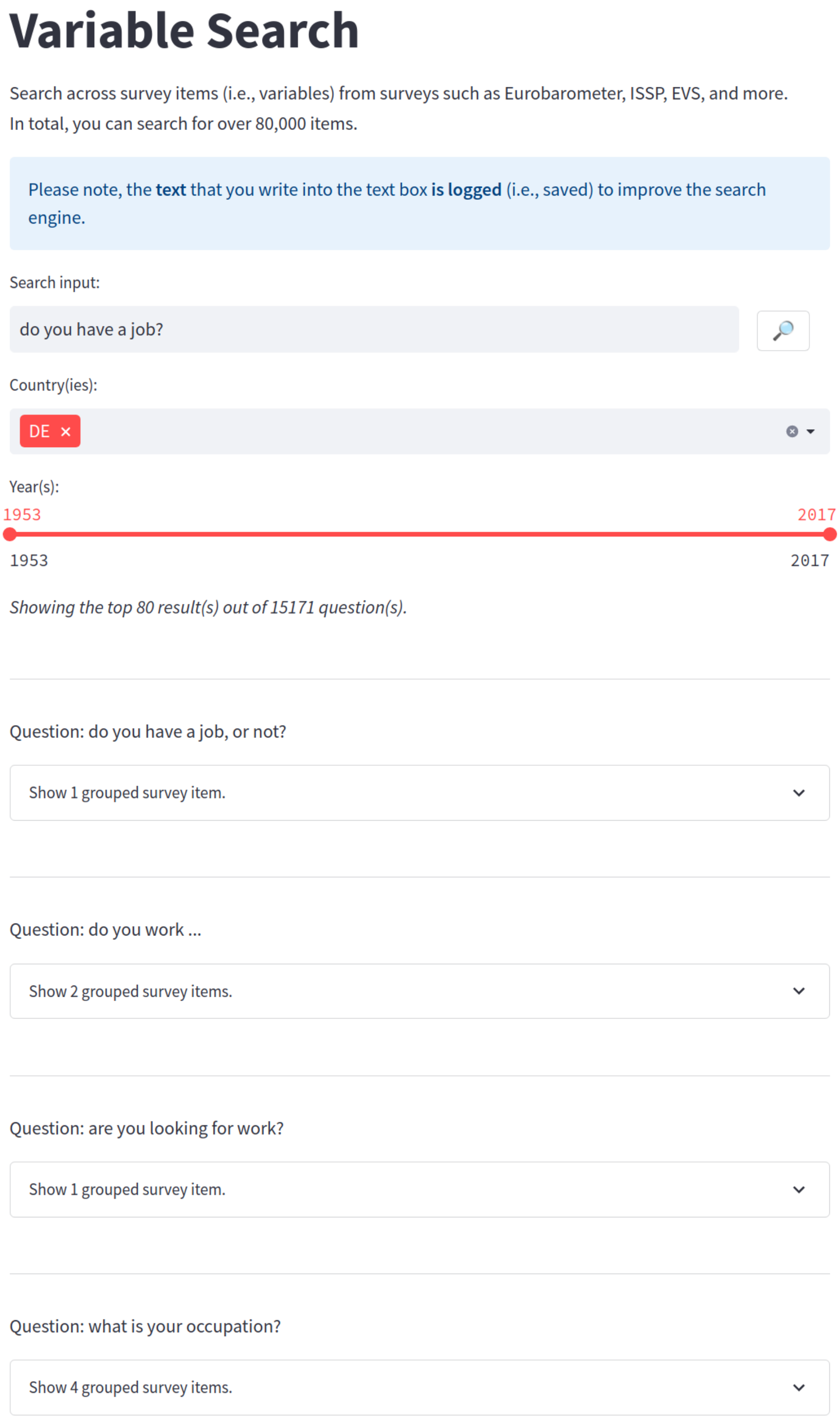
[2] Andrea Papenmeier, Thomas Krämer, Tanja Friedrich, Daniel Hienert, and Dagmar Kern. Genuine information needs of social scientists looking for data. *Proceedings of the Association for Information Science and Technology*, 58(1):292–302, 2021.

[3] Peter Granda, Christof Wolf, and Reto Hadorn. *Harmonizing Survey Data*, chapter 17, pages 315–332. John Wiley & Sons, Ltd, 2010.

[4] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022.

[5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[6] Robert Litschko, Ivan Vulic, Simone Paolo Ponzetto, and Goran Glavas. On cross-lingual retrieval with multilingual text encoders. *Inf. Retr. J.*, 25(2):149–183, 2022.



Variable Search

Search across survey items (i.e., variables) from surveys such as Eurobarometer, ISSP, EVS, and more.
In total, you can search for over 80,000 items.

Please note, the **text** that you write into the text box **is logged** (i.e., saved) to improve the search engine.

Search input:

do you have a job?

Country(ies):

DE ×

Year(s):

1953 2017

1953 2017

Showing the top 80 result(s) out of 15171 question(s).

Question: do you have a job, or not?

Show 1 grouped survey item.

Question: do you work ...

Show 2 grouped survey items.

Question: are you looking for work?

Show 1 grouped survey item.

Question: what is your occupation?

Show 4 grouped survey items.

Figure 1: Screenshot of the user interface of our search system for survey variable search.