Toward automatic preprocessing of complex free response data

Jordan Gunn

This paper explores the potential of automatic natural language processing (NLP) techniques to preprocess complex free response data in a range of domains. We examine various frameworks that break down the task of coding free response data into a series of discrete preprocessing steps and evaluate various automated techniques for each step. Specifically, we compare human and computational methods for segmenting text, matching response units with target items, and identifying the sequence of target items generated in a trial. We conclude with a discussion of the challenges and opportunities for automated free response data analysis and identify future directions for research in this area.

Table of contents

1 Introduction **4**
1.1 The potential of NLP for automating preprocessing 5
1.2 Research question and objectives 6

1 Introduction

Free response data consist of unstructured, open-ended answers generated by individuals responding to prompts or cues. Unlike structured data, collected through closed-ended probes that provide limited response options, research participants produce free response data in their own words, enabling additional insight into their thought processes and mental representations. In research across various fields such as psychology, sociology, and education, free response data is often analyzed qualitatively to understand participants' complex and varied experiences, attitudes, and perspectives (Hollway & Jefferson, 2000). In cognitive science, qualitative research methods are comparatively rare, but free response data is nonetheless central to the study of the mental processes underlying human behavior. Quantitative analyses of the ordering of freely generated responses and the time taken to produce them have provided key constraints for accounts of how people retrieve information from memory, make decisions, and solve problems (Ericsson, 2006; Kahana, 2020).

Free response data has been particularly influential in the study of memory search. For example, in the free recall task, participants are asked to remember a list of studied items in the order they come to mind. Data from the free recall paradigm reliably exhibits a temporal contiguity effect, in which items studied near one another are more likely to be associated and retrieved near one another. Experiments and analyses confirming this effect's time-scale invariance, automaticity, and forward-asymmetry are central to ongoing debates about the mechanisms underlying search through episodic memory (Healey et al., 2019). Similarly, in the semantic fluency task, participants are prompted to generate as many exemplars as possible that belong to a specified category. The order of responses in the semantic fluency task has fueled debate about whether semantic memory is typically searched through a random walk, optimal foraging, or another process (Kumar, 2021).

For free response data in this and many other domains, initial steps for preprocessing free response data to support organizational analysis can be broken down into solving two problems: (1) segmenting the response into discrete units of text ("response units") and (2) matching these units to the set of target items or features being researched. In the free recall task, for example, the text of a participant's free recall response is segmented into units corresponding to the words or phrases that participants generate in response to the prompt. In the semantic fluency task, the text of a participant's response is segmented into units of text corresponding to the words or phrases that participants generate in response to the prompt. In both cases, the units of text are then matched to elements in a set of target items; these are defined as the items studied in the encoding phase in the free recall task and as the pool of exemplars or features relevant to a cue in the semantic fluency task. Response units that cannot be matched

to target items are discarded or considered intrusions. The result for each set of responses is a sequence of target items considered to be generated by participants in the order they appear in the sequence.

While coding data in this way is straightforward enough for preprocessing small datasets, these demands forestall the use of free response paradigms in large-scale studies involving large quantities of participants and designs involving many trials per participant. Additionally, manual segmentation and correspondence can introduce bias and inconsistencies into the data, leading to inaccurate results. Preprocessing is all the more challenging in free response tasks involving the production of narratives, concepts, or autobiographical accounts where structured multi-word responses constitute the unit of scientific interest. Complex free response data such as this require sensitivity to grammatical conventions and semantic content to appropriately segment into ordered response units and correspond with target items. Standards for coding such data are necessarily vague, often requiring multiple reviewers to preprocess the same data samples to confirm interrater reliability. Even when these measures are taken, subtle differences in rater interpretation of these standards contribute to research degrees of freedom that can make it harder to interpret results and compare findings across studies (Simmons et al., 2016). An increasing emphasis on organizational analyses in memory research and other domains makes it especially urgent to address these limitations and develop methods for more efficient preprocessing of free response data.

TODO: Above needs citations documenting the obstruction and the importance of clearing it, probably findable by locating other efforts at automating coding.

1.1 The potential of NLP for automating preprocessing

Over much of the work using free response data, these problems have mainly been insurmountable, with researchers either relying on manual coding of free response data or, especially for larger-scale datasets, focusing on more structured paradigms. However, a cascade of technological advances in automatic natural language processing (NLP) has raised the possibility that many aspects of the work involved in preprocessing complex free response data can be automated, with human raters perhaps focused on confirming or correcting outputs. NLP is a subfield of artificial intelligence that focuses on developing computational methods for processing and analyzing human language. NLP techniques have been used to automate various tasks in related domains, including sentiment analysis, topic modeling, and machine translation. For example, in customer feedback analysis, NLP can help organizations quickly identify common themes and sentiments from large amounts of customer feedback data. In social media analysis, NLP can help identify emerging trends and sentiments related to a particular topic. In chatbot interactions, NLP can help generate appropriate responses to a user's questions and requests.

When these technologies work well, NLP has several advantages over manual methods for preprocessing free response data.

Techniques leveraging this technology have already been proposed for preprocessing less complex outputs, such as a free recall of word lists (for example, see). However, manual coding remains the standard approach to preprocessing complex free response data as ordered sequences of generated target units suitable for organizational analysis.

A range of NLP techniques has been developed that may be suitable for the two core problems of preprocessing complex free response data for organizational analysis. When it comes to segmenting complex discourse structures (rather than sequences of semantically independent words),

TODO: Above needs a review of significant examples of work attempting to automate the coding of both straightforward and more complex free-response data and outline the domains in NLP applicable to this task

1.2 Research question and objectives

In this paper, we examine the suitability of NLP for preprocessing complexly structured free response data and provide a brief overview of state of the art in automated free response data analysis. We develop a framework that decomposes the task of coding complex free response data into a sequence of discrete preprocessing steps and identifies techniques that may automate these activities. Focusing on a few especially promising approaches, we evaluate these techniques for data pre-processing across a range of free-response datasets, including free recall and word sense semantic fluency. We separately compare human and computational methods for segmenting free response data into units of text, corresponding response units with target items, and identifying the ordered sequence of target items generated in a trial. We conclude by discussing challenges and opportunities for automated free response data analysis and identifying areas for future research.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge Handbook of Expertise and Expert Performance*, 223–241.

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*(3), 699–720.

Hollway, W., & Jefferson, T. (2000). *Doing qualitative research differently: Free association, narrative and the interview method*. Sage.

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*, 107–138.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*, 40–80.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2016). *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.*