Transfer Learning for Auto-Coding Free-Text Survey Responses

Peter Baumgartner; Amanda Smith; Murrey Olmsted; Dawn Ohse; Bucky Fairfax

FedCASIC 2021

Research Question

Can responses to an open-ended survey question be accurately and automatically coded with machine learning?

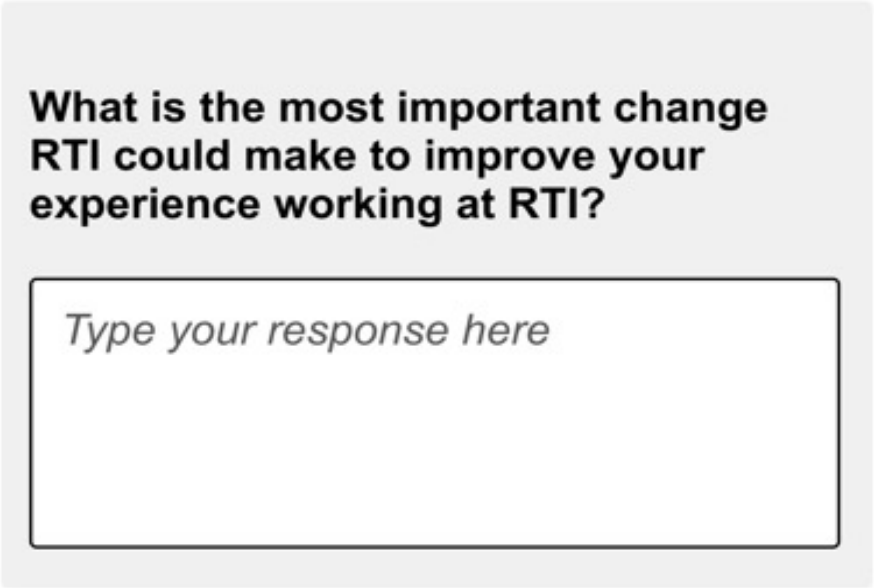
RTI Internal Employee Survey

- Administered to more than 4,500 RTI employees in 2018 and 2019
- Contains primarily the same items every year
- Used for action planning by leadership

Includes the open-ended question:

What is the most important change RTI could make to improve your experience working at RTI?

Open-Ended Questions – Why include them?

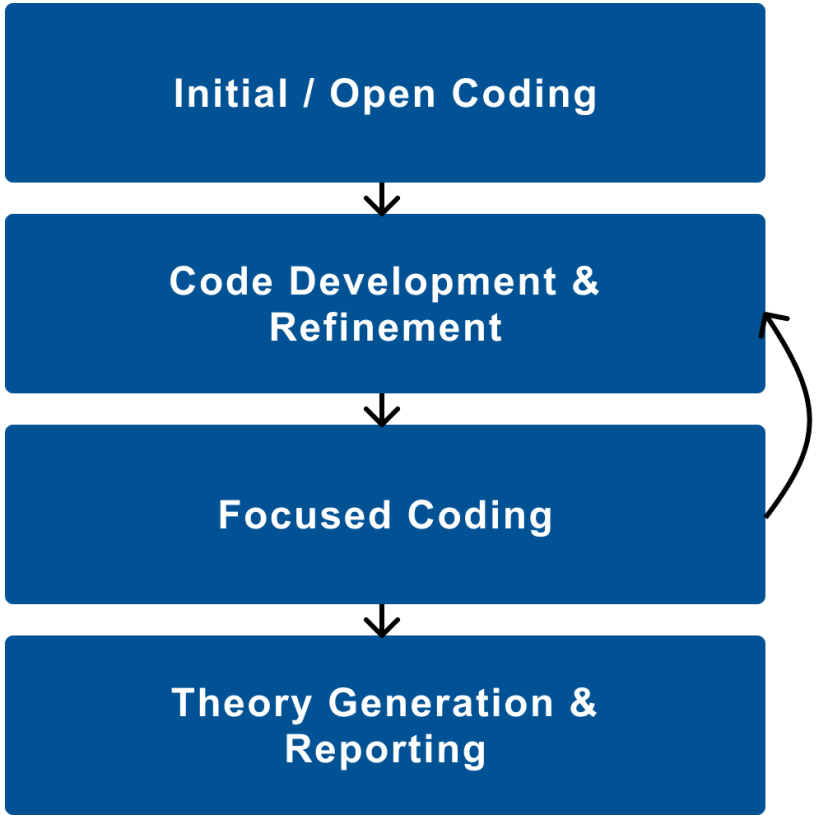


- Add depth and nuance to quantitative findings
- Can identify new information about attitudes and opinions
- Provide additional understanding of phenomena for development of future quantitative measures

Qualitative Coding of Open-Ended Questions

What is the most important change RTI could make to improve your experience working at RTI?

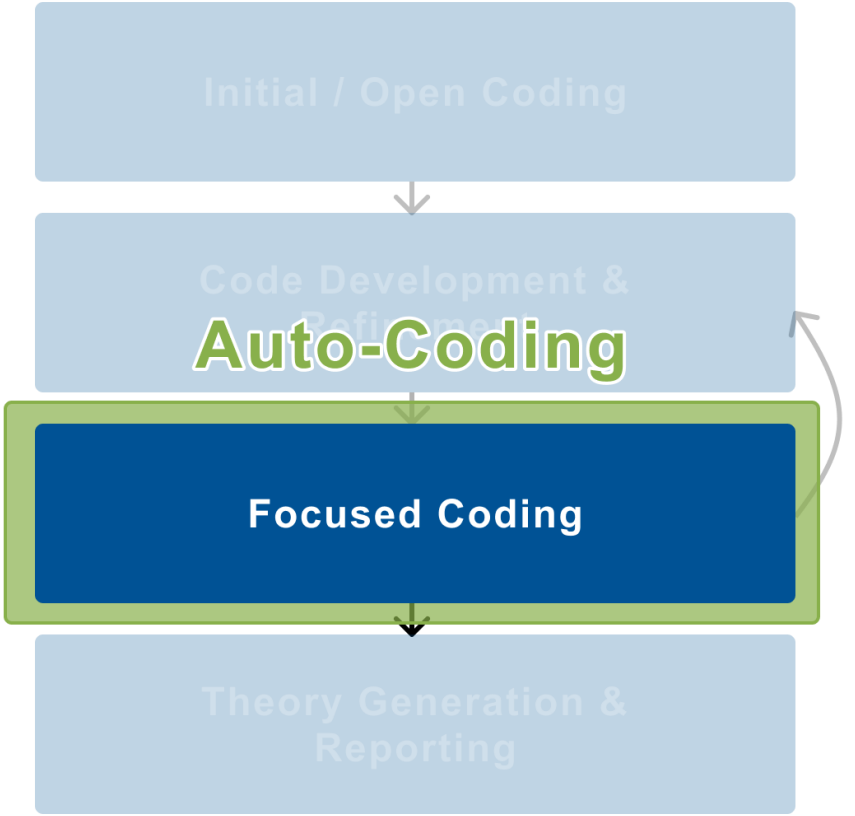
Type your response here



Where Auto-Coding Occurs

What is the most important change RTI could make to improve your experience working at RTI?

Type your response here



BERT & Transfer Learning

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

2018

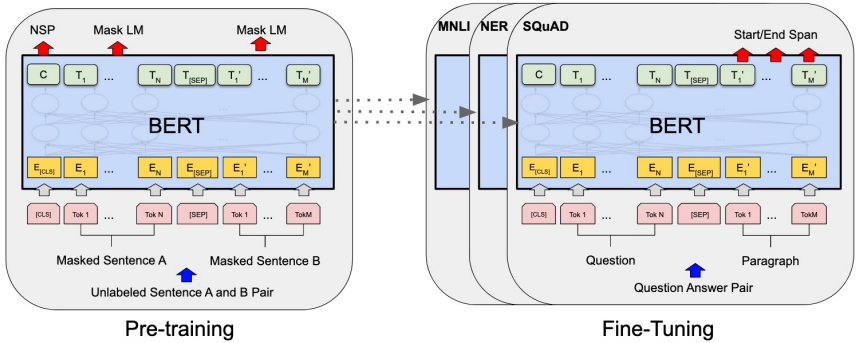
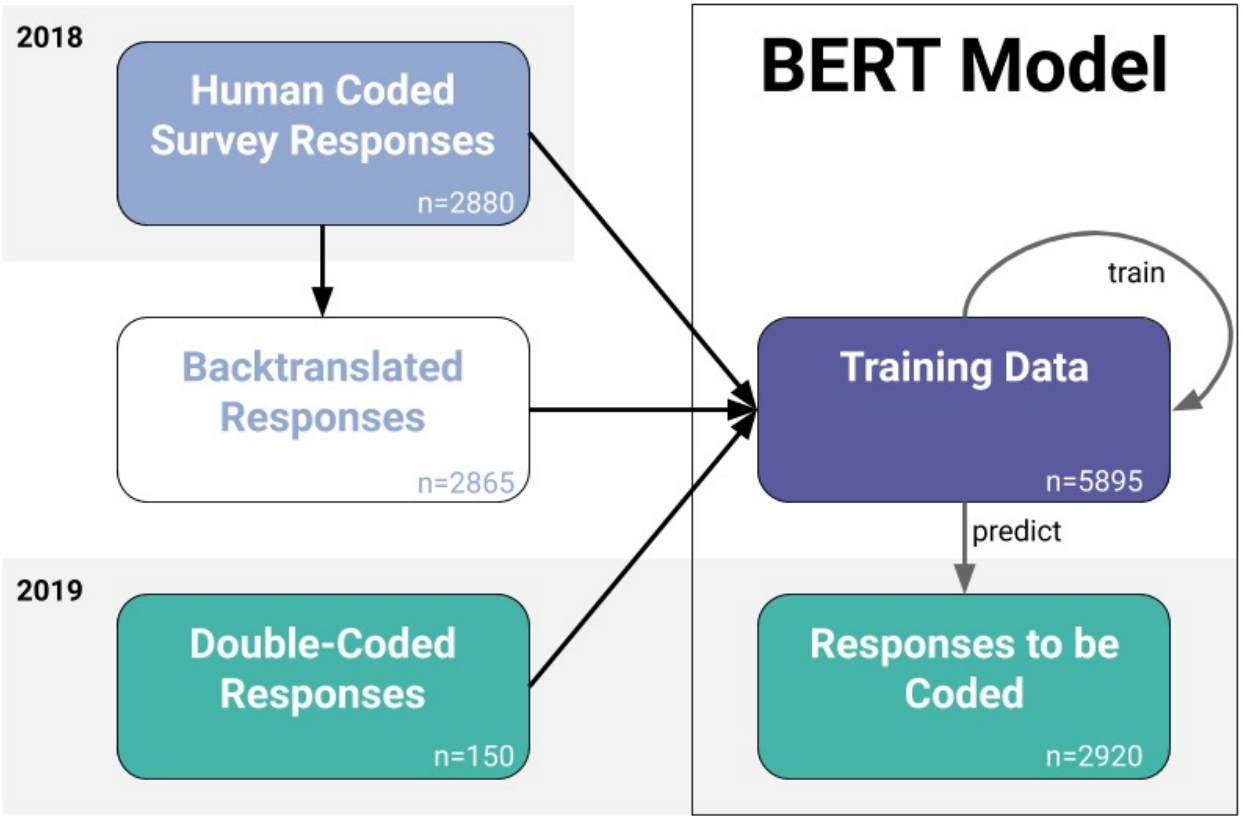


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Model Data Pipeline



Predictions Data

Response	Code 1	Code 2	Code 3	Code 4
We need more professional development opportunities	x	x		
RTI is doing great! I love it here				x
My manager has been too busy to support my development		x		

Predictions Data

Response	Code 1	Code 2	Code 3	Code 4
We need more professional development opportunities	x	x		
RTI is doing great! I love it here				x
My manager has been too busy to support my development		x		

False Positive

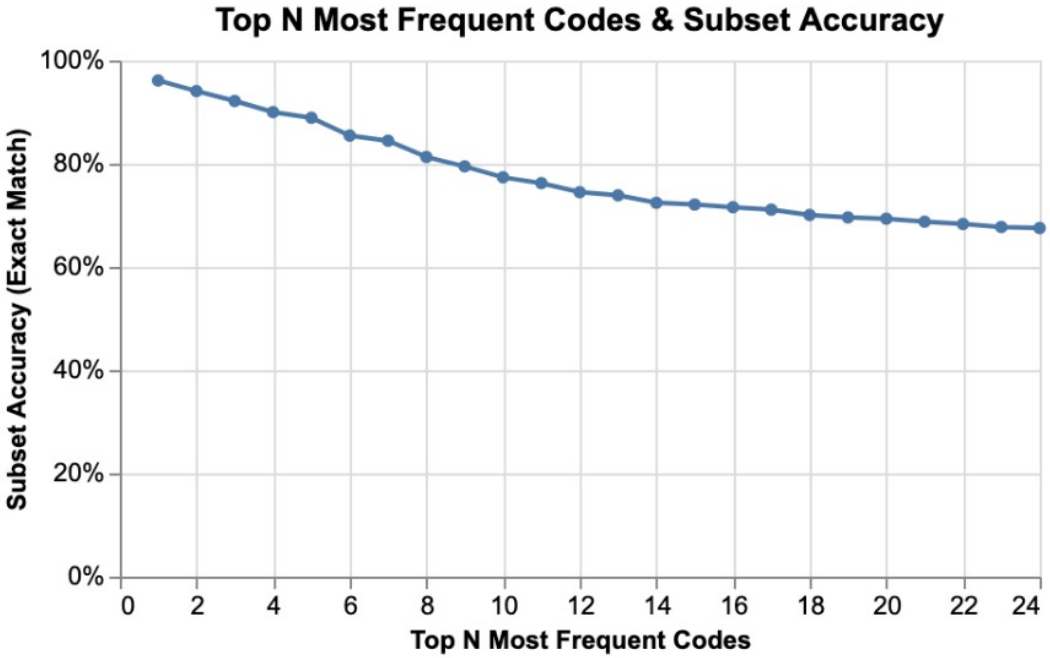
False Negative

Evaluating Performance: Overall Measures

Subset Accuracy or **Exact Match**:
the percentage of responses where the set of predicted codes exactly matched the set of codes after human review

Hamming Loss:
The percentage of code predictions that required a correction after human review.

Overall Performance



67.5% Exact match (all)

88.9% Exact match (top 5)

1.8% Hamming Loss

Benefits

- Simpler and more efficient task
  - Confirming codes instead of applying codes
- Speeds up the coding process, allowing results to be acted on sooner
- Consistency (i.e., no issues with intercoder reliability)

Limitations

- Requires sufficient manually coded data for model training
- Can’t identify new codes
- Lower performance for less frequent codes
- Requires specialized computational resources

Considerations for Use

- Longitudinal or repeated surveys that consistently ask an open-ended question
- Surveys where responses have already been coded

Thank you

Peter Baumgartner
Research Data Scientist
RTI International

pbaumgartner@rti.org