IRS

Enora Rice, Pranay Varada

Tagging IRS.gov Webpages with the BERT Language Model

08/19/2022

A Bit About Me

- Civic Digital Fellow with Coding it Forward
- First-Year PhD at the University of Colorado Boulder
- Studying computational linguistics

A Bit About Me

- Civic Digital Fellow with Coding it Forward
- Sophomore at Harvard College
- Studying Statistics

Research Question:

How informative are the text representations generated by taxBERT compared to a general model?

History of BERT

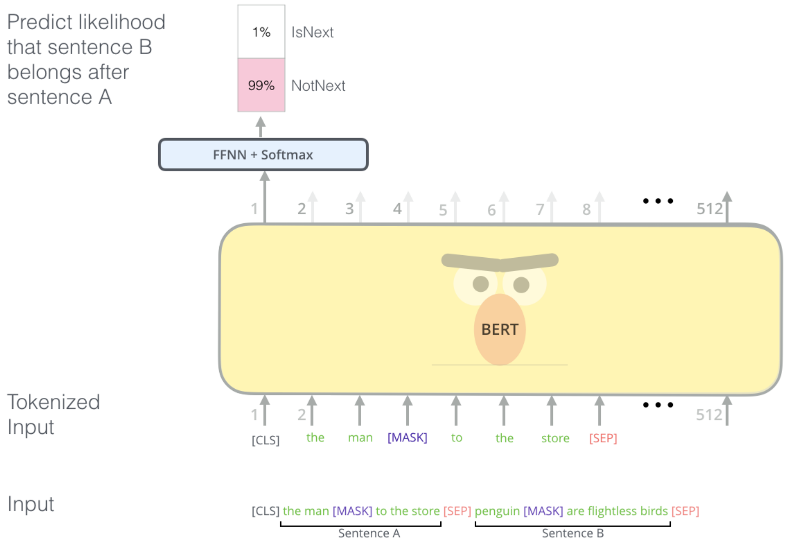
BERT: Bidirectional Encoder Representations (from) Transformers

- **Released by Google in 2018**
- **Pretrained on ~2.5B words/sentences on Wikipedia**
- **General model -- can then be further ‘tuned’ to your purposes**
- **Our project: comparing BERT with taxBERT**

Technical Overview

BERT Architecture

- **Takes in sentences, learns a “vocabulary” of words/sub-words**
- **Mathematically represents “meaning” of each word/sub-word**
- **Incorporates “context” (“river bank” vs. “bank account”)**


Predict likelihood that sentence B belongs after sentence A
1%
IsNext
99%
NotNext
FFNN + Softmax
1
2
3
4
5
6
7
8
...
512
BERT
Tokenized Input
[CLS]
the
man
[MASK]
to
the
store
[SEP]
Input
[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]
Sentence A
Sentence B

Pretraining versus Finetuning

There are two ways BERT can be adapted for IRS tasks

- Pretraining: the general BERT model can be further trained on domain specific corpus
  - Unsupervised, BERT learns through next token prediction
- Finetuning: task-specific machine learning models built off of a BERT base
  - Supervised, requires labeled training data
  - Can be used for classification

Prior RAAS work with NLP

“taxBERT”

- **RAAS has previously trained English BERT models on millions of sentences from publicly available documents**
  - IRC, IRM, form instructions, form notes, other publications
- **One use for IRS: answering FAQs via chatbot**
- **Classification of webpages that contain useful information can help with this task**

What tasks can BERT be tuned for?

Classification, Prediction Tasks

- **Sentiment analysis: positive/negative?**
- **Word prediction: What word fits best?**
- **Document summarization**
- **Question answering**
- **Text classification**

Scoping the Project

Data Wrangling, Code Adapting, Learning

- **First: did background research on projects leveraging taxmBERT and Spanish call transcripts—translation layer, sentiment analysis, topic modeling**
  - Found that BERT is not the best model for translation, topic modeling seemed most promising
  - Originally, we were focused on the new multilingual taxmBERT model developed by MITRE
    - Environment constraints lead us away from working with this model
  - Ultimately decided to switch to the English-only taxBERT model and focus on classification due to time constraints

Main Challenges

Working around Limitations

- **Consistently searching for new environments**
- **Project scope fluctuated constantly as new unforeseen barriers emerged**
- **Analytical environment down for a week due to recertification requirement**

Project Goal

Classify IRS.gov webpages using a predetermined list of tags

About IRS	Affordable Care Act	Business Tax
Charities and Non-Profits	Credits	Compliance
Deductions	Filing	Fraud and Scam
Government Entities	Income Taxes	International Taxes
Refund	Tax Professionals	

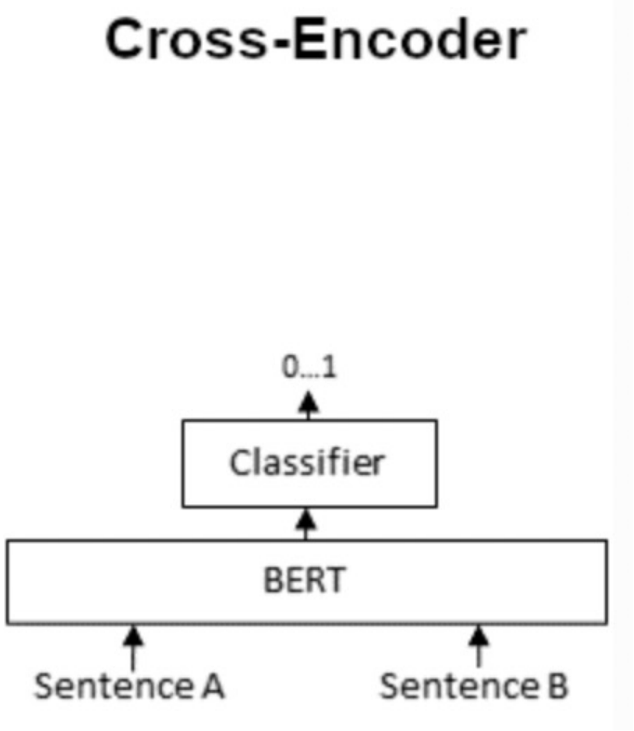
Purpose:

- **Cost-effective automatic labeling**
- **Metadata for search engines, dynamic webpages**

How We Approached Classification

Training a classification layer requires a large amount of labeled data

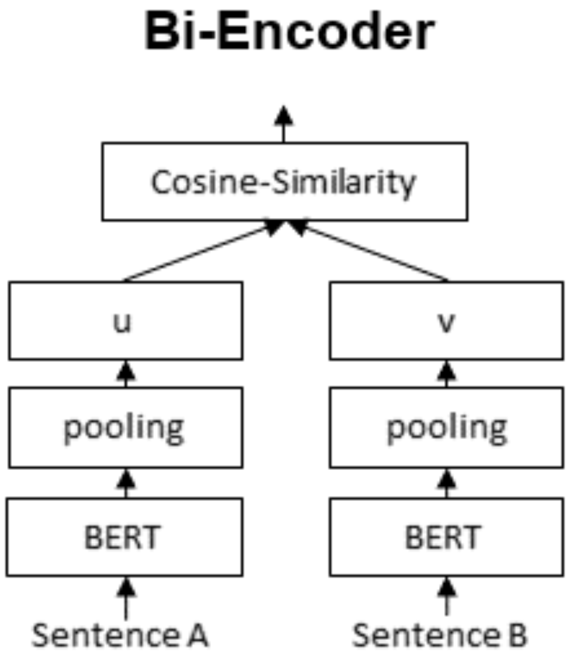
- Due to time constraints, we could not generate sufficient labeled data



How We Approached Classification

Settled on a bi-encoder approach to assess semantic equivalence instead

We can use BERT as an encoder without tuning it to a specific task—this allows us to generate vector representations of text that we can compare to one another



Results

Better than random, but not ready

- **In samples, taxBERT classified ~50% correctly, GenEng ~20%**
- **Why?**
  - Looking under the hood, it appeared that both models found the similarity of each text with each tag was ~0.4 - ~0.6
  - It seems that both models understood on a basic level that both the tags and the texts were related to the same topic – taxes – but not much nuance beyond that
  - Models even mislabeled texts directly included one of labels frequently

Takeaways and Future Work

Expanding taxBERT’s capabilities

TaxBERT shows promise for webpage tagging but is not sufficient

- Likely because taxBERT was trained on "legalese" documents rather than plain English

Possible Extensions

- Further pre-train a model on webpage data
- Recommended: manually assemble a corpus of labeled webpages to finetune a classifier
  - Would require labor upfront but ultimately less than manually labeling the whole corpus

Reflecting on the Summer

Working at the IRS

- **Often difficult to move past permissions, security measures, and interruptions**
- **Learning curve working with both old and new technology**
- **Great to see that RAAS DMD is a big part of the commissioner’s and the IRS’s vision for the future**
- **Hopeful that this summer’s work can be a step towards that vision**

IRS

Thanks for listening!