Kedar Garzón Gupta

taxBETO: Training a Spanish Tax Domain-Specific Language Model

8/12/21

Research Question:

How can we build a machine model that effectively represents Tax Content in Spanish?

History of BERT

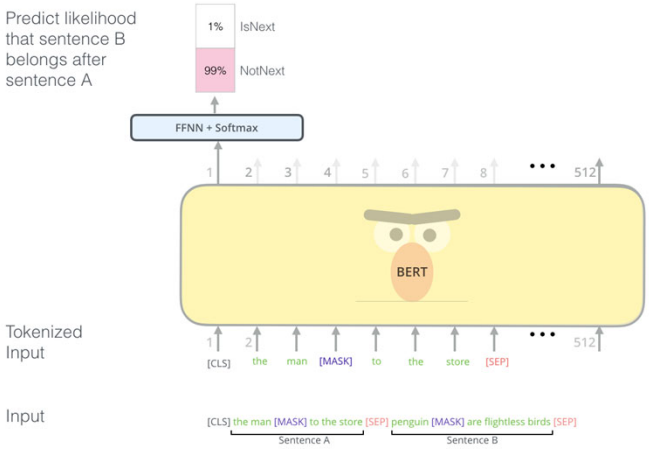
BERT: Bidirectional Encoder Representations (from) Transformers

- **Google released BERT in 2018**
- **Pretrained on ~2.5B words/sentences on Wikipedia**
- **General model -- can then be further 'tuned' to your purposes**
- **My project: training domain-specific model specifically on Spanish tax content**

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”, “bank account”)**



What tasks can BERT be tuned for?

Classification, Prediction Tasks

- **Sentiment Analysis: Positive/Negative?**
- **Word Prediction: What word fits best?**
- **Semantic Equivalence: Sentence A == Sentence B?**

- ***Of special interest to RAAS:**
- **Machine Translation**
- **Question Answering**

Prior RAAS work with NLP

"taxBERT"

- **RAAS has previously trained English BERT models on millions of sentences from irs.gov + form PDFs**
- **One use for IRS: use BERT to answer FAQs via chatbot**
- **Spanish is most-spoken language for non-English speakers**
- **So, multilingual BERT can assist Spanish-speaking taxpayers via chatbot**

What did I do?

Data Wrangling, Code Adapting, Learning

- **Researched pre-trained Spanish BERT models, found “BETO” by University of Chile (trained on Spanish Wikipedia)**
- **Aggregated data from various sources**
- **Adapted code from existing taxBERT work to use BETO**
- **Trained models over course of several days, ran benchmarks to assess performance**

Main Challenges

Working around Limitations

- **Can’t run scripts that download files from the Internet – interesting challenge**
- **Can only use Spanish sentences – had to ignore manually-translated Spanish phrases from IRS corpora**
- **BERT is a complex architecture – took a while to ‘train’ myself on its inner workings**

Results: Model Performance

Task (Corpus)	taxBETO	BETO-uncased	Best multiBERT
Entailment (XNLI)	69.75	80.15	78.50
Named Entity Recognition (CoNLL)	79.10	82.67	87.38
Semantic Equivalence (PAWS-X)	86.88	89.55	90.70

Future: developing a set of prediction tasks and comparing originalBETO to taxBETO.

Demo: Prediction Tasks

#1: “Alguien puede [reclamar] a usted como un dependiente para recibir [deducciones]”
TR: Someone can [claim] you as a dependent to receive [deductions]

taxBETO:
MASK 0 : [**'reclamar'**, 'tratar', 'usar', 'usarlo', 'incluir’] (**claim**, treat, use, use it, include)
MASK 1 : ['##lo', '##la', 'ayuda', **'reembolso'**, 'pagos’] (recibirlo/la = “receive it”, help, **reimbursement**, payments)

originalBETO:
MASK 0 : ['venir', 'ser', 'referirse', 'ir', 'dirigirse’] (come, be, refer, go, direct oneself)
MASK 1 : ['[UNK]', **'dinero'**, 'ordenes', 'ayuda', 'algo’] (unknown, **money**, orders, help, something)

#2: “Free File del IRS le permite preparar y presentar su declaración de [impuestos] federales sobre los ingresos gratuitamente..”
TR: IRS Free File lets you prepare and present your federal [tax] declaration concerning your income, for free

taxBETO:
MASK 0 : [**'impuestos'**, 'impuesto', 'contribuciones', 'declaraciones', '[UNK]’] (**taxes**, tax, contributions, declarations, unknown)

originalBETO:
MASK 0 : [**'impuestos'**, 'derechos', 'los', 'impuesto', 'datos’] (**taxes**, rights, the, tax, facts)

Where does this leave us?

So what?

- **Q: Is this model ready to be effectively used for translation/chatbot purposes?**
- **A: No, performance is not quite there – but having this research highlights the urgent need for Spanish data collection efforts.**

- **Q: Where can we get more Spanish data?**
- **A: The challenge is needing Spanish sentences about U.S. Tax Law; if we feed e.g. Spanish accounting textbooks.**

- ***Room for optimism that the necessary data can be found.**

Reflecting on the Summer

Working in Government

- **Challenging: often working around regulations, security measures**
- **Meaningful: feels like work will genuinely help Americans downstream**
- **Surprising: IRS is quite technologically advanced, in terms of faculty/knowledge base**

Thanks for listening!