Estimating Search Relevance using Modern Deep Neural Networks

Semantic Search Relevance

Aditi Duggal
CS-584: Natural Language Processing
May 3rd, 2022

Introduction

- As more and more people shop online, customers pay more attention to online shopping experience.
- Many bad shopping experiences come from the difficulty in finding the right products.
- If online retailers can more accurately predict the relevance between search terms and products and pop out the products that can better match customers' need, this is extremely attractive and interesting.
- Therefore, many online retailers are working on

  such a relevance model.

Problem Statement

- Given product information and search term, develop a model which can accurately predict the relevance score of a product.
- Relevance score?
  - The relevance score describes how relevant a product is to a given search term.
  - The relevance is a number between 1 (not relevant) to 3 (highly relevant). For example, a search for "AA battery" would be considered highly relevant to a pack of size AA batteries (relevance = 3), mildly relevant to a cordless drill battery (relevance = 2), and not relevant to a snow shovel (relevance = 1)

Data Overview

Dataset Link: https://www.kaggle.com/competitions/home-depot-product-search-relevance/data

- Source: Kaggle -Home Depot Product Search Relevance
- Dataset consists of 74K observations
- Files:
  - train.csv - the training set, contains products, searches, and relevance scores
  - product_descriptions.csv - contains a text description of each product.
  - attributes.csv - provides extended information about a subset of the products
- I used 80-20 train test split
  - Train size: 47402
  - Test set: 14814

- Data fields:
  - id - a unique Id field which represents a (search_term, product_uid) pair
  - product_uid - an id for the products
  - product_title - the product title
  - product_description - the text description of the product search_term - the search query
  - relevance - the average of the relevance ratings for a given id
  - name - an attribute name
  - value - the attribute's value

Data Pre-processing

- The given information about each product is somewhat poor-structured.
  - Appended product description to the dataset based on the Product ID
  - Extracted brand information from the attributes dataset and appended to the train dataset
  - Performed the following:
    - Converted all the words to lowercase
    - Eliminated punctuations
    - Tokenize
    - Removed stop words
    - Lemmitize the words

Data Pre-processing

Prepared data

```
df.head()
```

	id	product_uid	product_title	search_term	relevance
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00
1	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50
2	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 ...	deck over	3.00
3	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	rain shower head	2.33
4	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	shower only faucet	2.67

```
df.head()
```

	id	product_uid	product_title	search_term	relevance	preprocessed_title	product_description	preprocessed_description	product_attributes	preprocessed_attributes
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00	simpson strong tie 12 gauge angle	Not only do angles make joints stronger, they ...	angle make joint stronger also provide consist...	Bullet01 Versatile connector for various 90° c...	bullet01 versatile connector for various 90° c...
1	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50	simpson strong tie 12 gauge angle	Not only do angles make joints stronger, they ...	angle make joint stronger also provide consist...	Bullet01 Versatile connector for various 90° c...	bullet01 versatile connector for various 90° c...
2	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 ...	deck over	3.00	behr premium textured deckover 1 gal sc 141 tu...	BEHR Premium Textured DECKOVER is an innovativ...	behr premium textured deckover innovative soli...	Application Method Brush,Roller,Spray Assemble...	application method brush roller spray assemble...
3	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	rain shower head	2.33	delta vero 1 handle shower faucet trim kit chr...	Update your bathroom with the Delta Vero Singl...	update bathroom delta vero single handle showe...	Bath Faucet Type Combo Tub and Shower Built-in...	bath faucet type combo tub and shower built in...
4	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	shower only faucet	2.67	delta vero 1 handle shower faucet trim kit chr...	Update your bathroom with the Delta Vero Singl...	update bathroom delta vero single handle showe...	Bath Faucet Type Combo Tub and Shower Built-in...	bath faucet type combo tub and shower built in...

Feature Engineering

Extracted Features

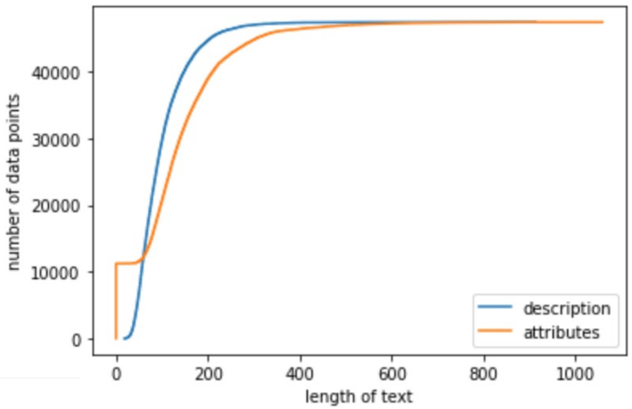
```
print("Maximum length of the titles: ",max_length_of_titles)
```

```
Maximum length of the titles:  32
```

```
print("Maximum length of the search term: ",max_length_of_search_term)
```

```
Maximum length of the search term:  17
```

Length of text vs # of data points




```
Vocab Lengths
title:  15143
description:  131825
attribute:  29024
search term:  7654
```

Evaluation Metric

Root Mean Square Error

- The results are evaluated on the root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i denotes the labeled relevance and $\hat{y}_i$ denotes the predicted relevance.

Model Training

Training methods

I will be exploring the following Neural Network models:

1. Bidirectional LSTM (**sequence processing model)**
2. One Direction Convolution (CNN)
3. Attention Mechanism
4. Transformers
    - BERT
    - Sentence BERT

Results

RMSE Values

MODELS	RMSE	MSE	MAE	TIME TAKEN TO PREDICT RELEVENCE
BiLSTM	0.488	0.238	0.3833	0.177
Conv 1D	0.529	0.281	0.377	16.28
Attention	0.529	0.280	0.385	4.18
BERT	0.528	0.579	0.384	11.50

*Couldn't complete the Sentence-BERT as it was a costly task

Analysis & Future work

- I have trained various neural network models including Bidirectional LSTM, One-dimensional Convolution, Attention Mechanism, and Transformer based models such as BERT to solve the problem
- Until now we got some good RMSE score of 0.488 from Bidirectional LSTM which has taken the least time of 0.177
- I believe with better GPU, we can achieve a Sentence-BERT which would give better results.
- With better feature extraction, we could achieve better RMSE score.
- We can display the top 10 or 20 products with higher relevancy score to the user. while building such search engines is not the scope of this project, we can use trained models to build such systems. Building such systems will help us display semantically similar products to the user.

References

- DHIVYA CHANDRASEKARAN and VIJAY MAGO, Evolution of Semantic Similarity arXiv:2004.13820v2 [cs.CL] 30 Jan 2021
- Data Link: https://www.kaggle.com/competitions/home-depot-product-search-relevance/data
- P. Zhou et al. “Attention-based bidirectional long short-term memory networks for relation classification,” in Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers) 2016, pp. 207–212.
- D. Tang, B. Qin, X. Feng, and T. Liu, “Effective LSTMs for target-dependent sentiment classification,” arXiv Prepr. arXiv1512.01100 2015.

STEVENS

INSTITUTE of TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu