Improving Information Retrieval for Knowledge Extraction & Open Book Question Answering

- Pratyay Banerjee, Kuntal Pal, Aditya Narayanan, Kunal Bagewadi, Prashanth Sukhapalli, Bhavani Balasubramanyam

Overview

1. Motivation
2. OpenBook QA - Why this work ?
3. Knowledge Extraction - Why is it hard ?
4. Our Approach - What is new ?
5. Knowledge Extraction - Methods, Challenges, Experiments
6. Question Answering - Methods, Challenges, Experiments
7. Insights - What did we learn ?
8. Conclusion and Future Works

Motivation

1. Information Retrieval (IR) and Knowledge Extraction(KE) is a core component for many NLP Tasks.

2. Especially in Open Domain Question Answering (QA), we need to choose among very similar facts. The facts need to be relevant and not redundant.

3. In this project, we improve on IR and KE for an application task of OpenBook QA. With improved KE, we also improve on OpenBook QA accuracy.

Our Contributions and Approach

- Deeper Natural Language Understanding: Evaluate BERT CLS Token Embedding and CNN based Sentence Embedding using BERT for IR and KE. Compare it against TF-IDF based features.
- Comparative empirical study of SML algorithms for the IR task.
- Choose best model and evaluate over OpenBookQA retrieval task.
- Evaluate a new BERT-QA model with CNN based Sentence Embedding.
- Understand the relevance of features present in CLS token using the observed feature importances.

OpenBook QA

- Dataset released by AllenAI. Contains 5957 Questions and each need additional knowledge to be able to answer the questions
- Dataset has a set of 1350 facts : The OpenBook, but not self contained.
- A complex QA task, requires multiple types of reasoning, such as multi-hop reasoning, conjunctive , temporal etc
- Most Importantly, the task requires precise Knowledge for systems to perform better

Question:

Which of these would let the most heat travel through?

A) a new pair of jeans.

B) a steel spoon in a cafeteria.

C) a cotton candy at a store.

D) a calvin klein cotton hat.

Science Fact:

Metal is a thermal conductor.

Knowledge Extraction

- To solve OpenBookQA , we need to do IR and KE over the OpenBook
- This is a challenging task as the science facts are very similar to each other, and requires Natural Language Understanding
- For ex:
  - Question : Beak shape can influence a bird's ability ?
  - Relevant Fact : Beak is related to food.
  - Distracting Fact : Beak and mate has no relation.
  - Answer Options : “to mate with it's partner” , “**to chew up certain worms**”

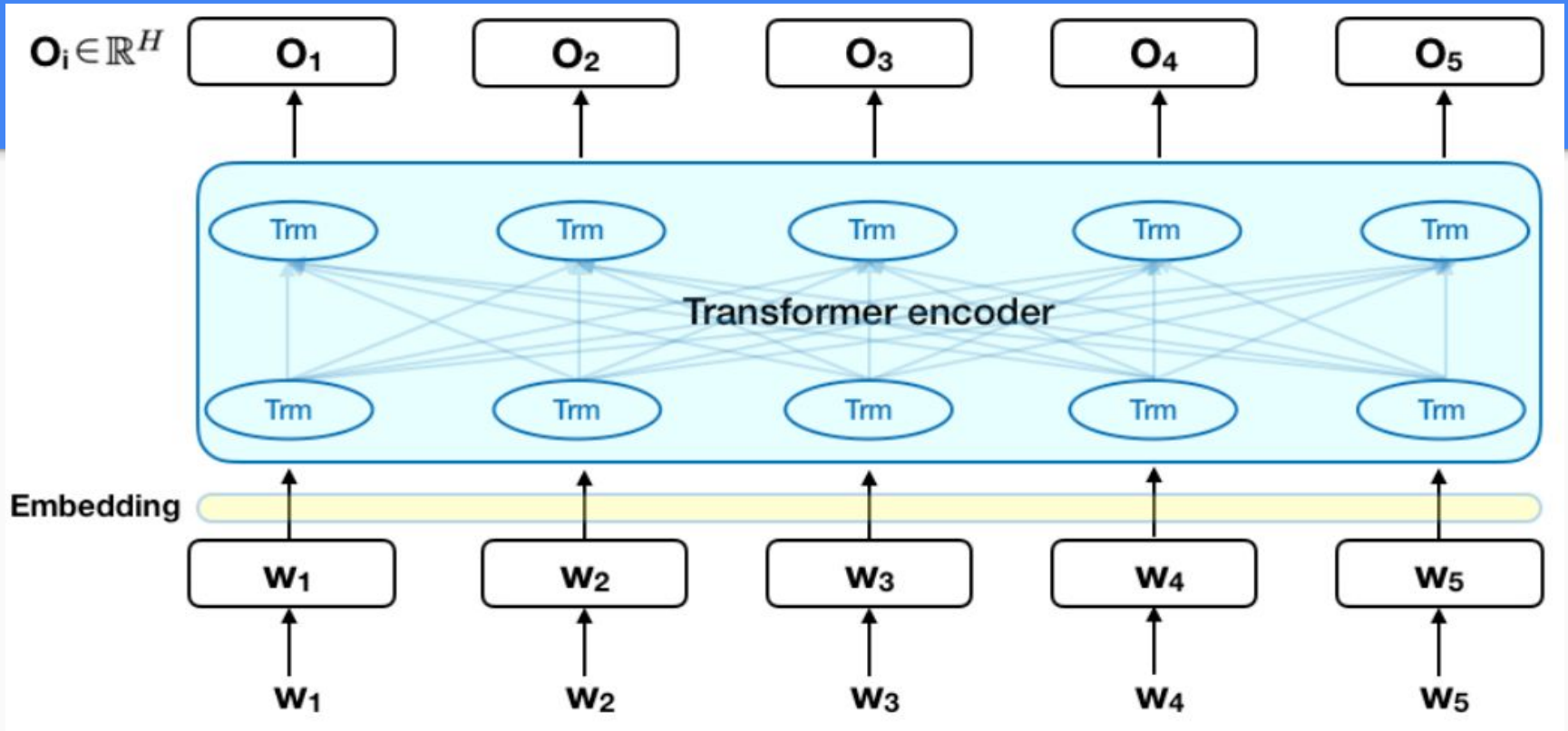
Dataset - Microsoft AI Challenge Dataset

- ‘Learning to Rank’ dataset , which has a similar domain as OpenBookQA.
- Original dataset: 5 Million Queries, each having 10 documents and 1 document labelled as the most relevant.
- Modified dataset:
  - Train Set : 75000 Query-Document Pairs
  - Validation and Test Set : 15000 each

Feature Generation - Using BERT

- **Baseline Feature : TF-IDF** scores for each Query-Document Pair
- For generating Sentence Embeddings, we **finetune BERT Base** using our entire corpus for few epochs. Both **LM** and **NextSentencePrediction**.
- **CLS Token Sentence Embedding** : We extract only the CLS token after feeding our Query-Document pair. Dimension : 768
- **CNN Based Sentence Embedding** : We concat last 4 layer's encodings of BERT Base including CLS token, feed it through a CNN . Dimension : 25344
- A **PCA based reduced Sentence Embedding** from the CNN generated embeddings. Dimension : 6000

Quick Glance of BERT



What do we Learn ?

- Relevance : - Score(Query,Document)
- How ? We have Labels : 0 for Irrelevant, 1 for Relevant
- How do we Rank? We take Probabilities for Class label 1, and sort them
- Do all Models do this ? Yes

Models and Challenges

- We evaluate all different Statistical ML algorithm families
- We also evaluate Deep Learning models with our input features
- We faced multiple system challenges for performing experiments, especially with CNN based Encodings due to size of input data.
- We resolved most of our issues by performing PCA and creating a smaller representation , with a cumulative sum of variance ratio being 90%. (Data retention is 90%).
- Dimensions were reduced from 25K to 6K for the CNN sentence embeddings.

Algorithms Evaluated For IR

- Linear Family
  - Logistic Regression
  - Passive Aggressive
  - Perceptron
  - SGD Classifier
- Tree Family
  - Decision Tree
  - Random Forest
  - Extra Tree
  - Extra Tree Ensemble
- SVM
  - Linear
  - Polynomial
  - RBF
  - Gaussian
- Naive Bayes
  - Gaussian
  - Multinomial
  - Bernoulli
- K-Nearest Neighbor
- Neural Network
  - 1-Layer
  - 2-Layer
- Boosting
  - GBM
  - XGBoost
- BERT
- BERT-CNN

Metrics

- **Accuracy** of Classifier - For the Binary Classification Task
- **Mean Reciprocal Rank** - Rest are for the IR Task
- **Precision@1** and **Precision@3** - Is not same as Accuracy. Classifier can classify all 10 documents as irrelevant, but still have a good Precision@1.
- Why only 1 and 3 ? Our final QA model is limited by how much knowledge it can take as input. Better 1 and 3 is the final goal.

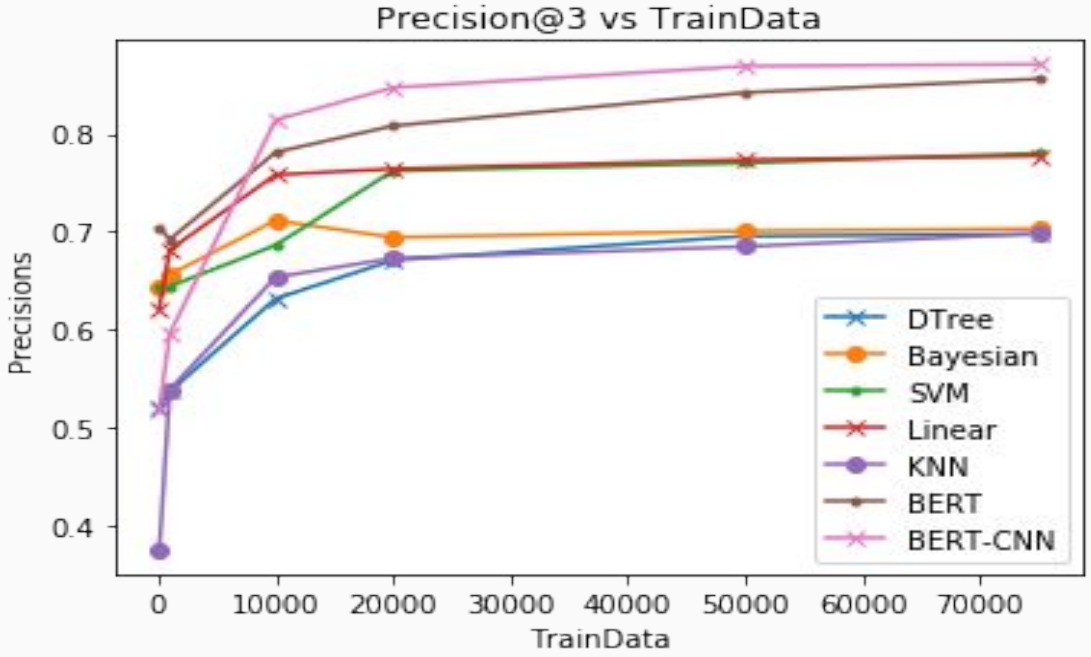
Evaluation - Comparison of Features

P3	*TFIDF*	*CLS*	*ALL*	*PCA*
SVM	66.86	78	NA	80
NN	**67.7**	80.1	80.4	**80.2**
BERT	NA	**83.8**	**84.8**	NA
DT	61	69	69	65
KNN	64.67	69.8	NA	63.4
Bayes	67.53	70.26	44.533	64.53
Linear	67.86	77.73	68.6	77.86

Evaluation - Comparison of Families

		CLS-Tokens	Validation 2		
	Best Model	*Accuracy*	*Mrr*	*P1*	*P3*
DT-Family	Random Forest	59.03	52.9	29.1	69.7
KNN	k=99	60.87	54.7	30	69.8
Naive Bayes	MultinomialNB	58.09	53.4	29.4	70.3
Linear Family	Logistic Regression	65.11	60.1	37.6	77.7
SVM	LinearSVC	65.89	61.6	39.4	80
NN	2 Layer NN	66.04	60.5	39.6	80.1
BERT	Base ,Batch:64, lr :1e-5	70.4	68.9	49.8	85.6
BERT-CNN	Base ,Batch:64, lr :5e-6	**71.2**	**69.6**	**50.3**	**86.1**

Learning Curves - CLS Token Features



Evaluation - Adding More Data & Params ?

Model - Data	Precision@3
BERT- Large 75K	0.873
BERT - CNN - Large 75K	0.881
BERT - Large 2M	0.904
BERT - CNN - Large 2M	***0.907***

OpenBookQA - Models

For the QA task, we evaluate 2 Models, each with retrieved knowledge. The Baseline model is the No Knowledge model.

- BERT - Large
- BERT - Large with Sentence Embeddings using CNN

We also evaluate the new IR model over OpenBook of the dataset. As Baseline we compare with 2 models, a TF-IDF model and a vanilla BERT Pretrained over STS-B dataset (A Textual Similarity Dataset).

Evaluation - IR over OpenBook

Model - Trained On Dataset	Precision@1	Precision@3
TF-IDF	228/500	324/500
BERT - STS-B	288/500	368/500
BERT - Large - MS AI	240/500	304/500
BERT - Large - OBQA	258/500	358/500
BERT CNN - MS AI	**314/500**	**389/500**
BERT CNN - OBQA	264/500	353/500

Evaluation - QA Task

Model (BERT Large)	Val Accuracy	Test Accuracy
No Knowledge - Leaderboard	60.2	60.4
Knowledge STS	66.8	66.2
BERT-CNN + Knowledge + MS AI	**67.4**	**67.2**
BERT-CNN + *Oracle* Knowledge	*74.2*	*74.4*

Insights

- BERT, a pretrained language model captures a lot of semantics and language structure, which is shown by the performance of other models using BERT features.
- Analyzing feature importances of CLS token encodings, lead to identification of only few dimensions being used by most models to determine effectively the Relevance. Indicating BERT captures semantics only a few dimensions.
- Using BERT encodings and a CNN as a feature increases semantic information and leads to better task performance, both IR and QA.
- Sentence embeddings and PCA of embeddings did not reduce classifier performance, enabled much more explorations and experiments.

Insights ..

- Most SML algorithms are not scalable for a huge dataset, whereas BERT was easily able to scale to 2M Query-Document pairs. SVM took 2 days to train with 75K pairs, BERT took 20 mins and with better performance.
- IR of OpenBook QA showed the Transfer Learning capability of BERT-CNN model trained over MS AI data.
- The improvement of QA task is only 7%, shows how hard the overall task is and more precise external knowledge is needed.

Insights ..

- KNN with 99 neighbours approached an acceptable performance, such a neighbourhood indicates a minimum neighbourhood for distinguishing Relevant and Irrelevant documents
- On the other hand, the SML models fail in comparison to fine-tuned BERT as expected, as features for the models come from an unsupervised BERT model, and a supervised BERT model will update its parameters to the required task and Domain much better.

Conclusion and Future Work

- Our hypothesis of using BERT features as Semantic information for better IR and KE was proved, but it's still not perfect.
- Our architecture of CNN over BERT encodings also improved IR and QA task, far from human accuracy.
- Analysis of the QA task, robustness and adversarial tests are future work.
- We explored the explainability of BERT features using feature importances and were able to get few insights. Can we establish a relation from CLS token features to concrete NL semantics and sense ?