TextGraphs 2019 Shared Task

Language Model Assisted Explanation Generation

RED DRAGON AI

EMNLP IJCNLP 2019

Yew Ken Chia ken@RedDragon.ai

Sam Witteveen sam@RedDragon.ai

Martin Andrews martin@RedDragon.ai

Summary

Shared Task :

- Rank explanation sentences for elementary school science questions

Data Used :

- WorldTree Corpus
- 'Common Sense' embedded in BERT

Ideas :

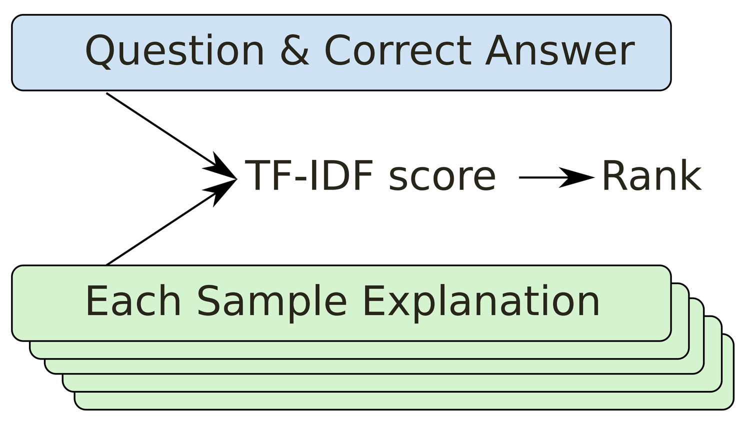
- Baseline TFIDF can be improved
- Iteratively 'grow' explanation set
- Use BERT to learn to rank explanations

Results :

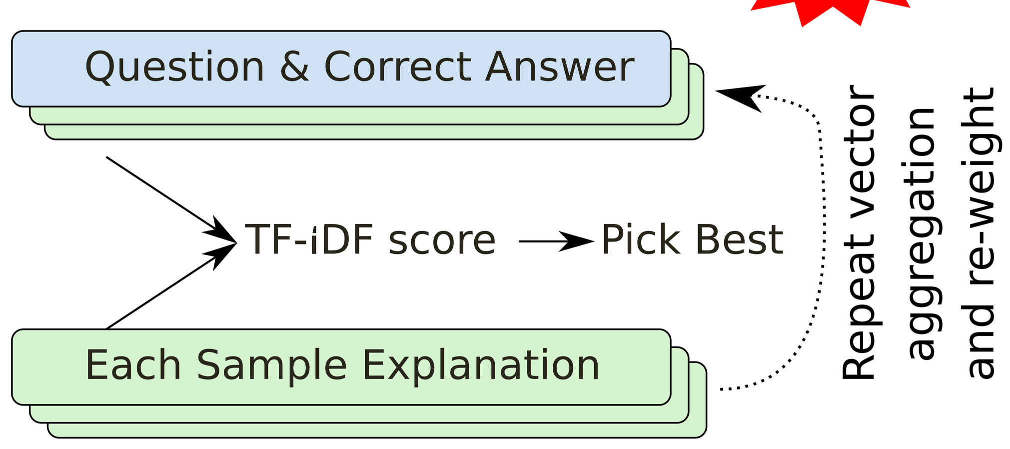
- Submitted score : 0.4017
- 3 better methods outlined here
- TF-IDF can take us a long way
- Final BERT method is learned

Three New Methods with Increasing Test Scores

Method 1 : Better TF-IDF Ranking 0.4274

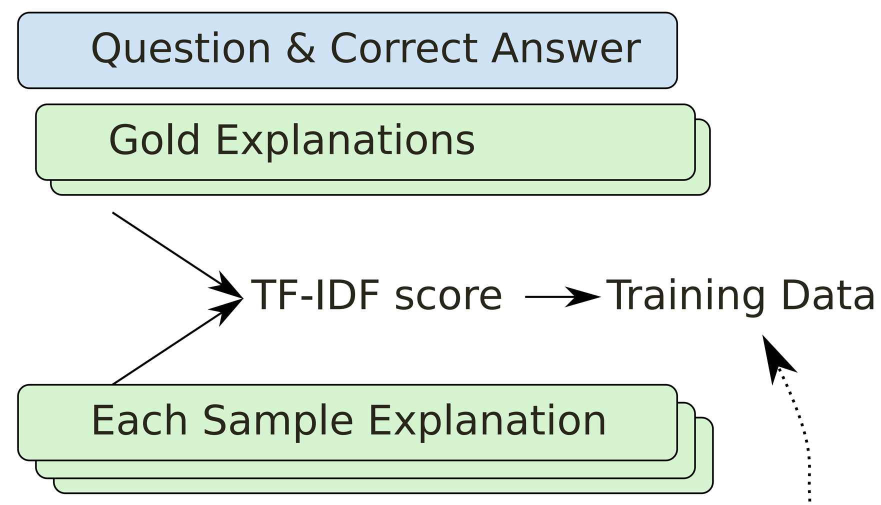


Method 2 : Iterated TF-IDF 0.4576

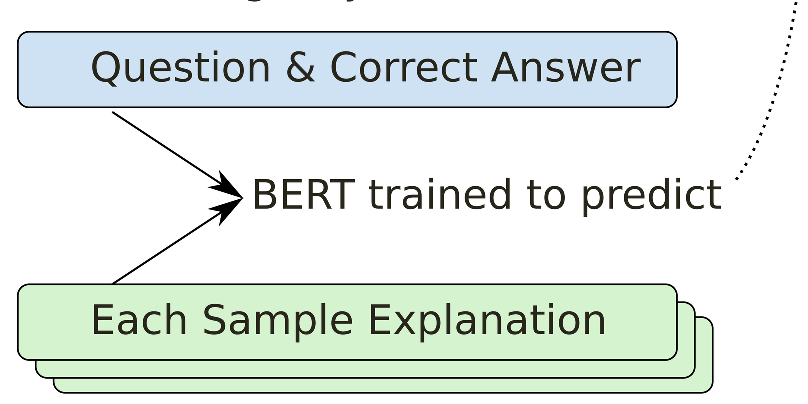


Method 3 : BERT predicts relevance 0.4771

Determine TF-IDF score targets using Gold explanations



Train BERT to predict 'informed' score above, using only available data



Results : Baseline, Submitted and New Methods

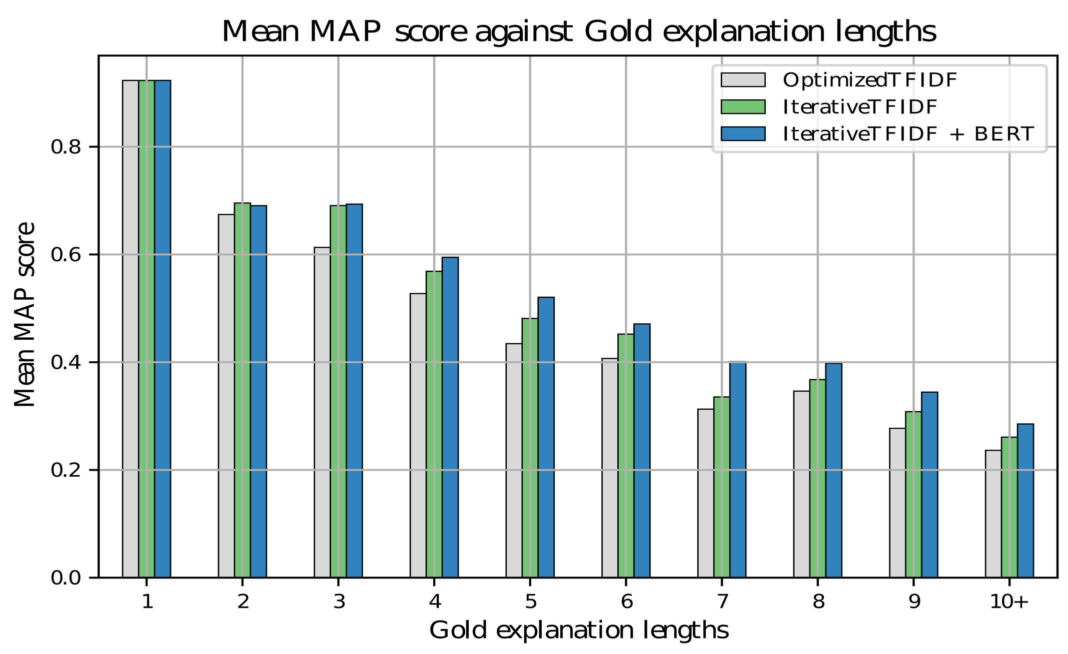
Data split	Python Baseline	Scala Baseline	Python Baseline^{1e9}	Leaderboard Submission
Train	0.0810		0.2214	0.4216
Dev	0.0544	0.2890	0.2140	0.4358
Test				0.4017

Table 1: Base MAP scoring - where the Python Baseline^{1e9} is the same as the original Python Baseline, but with the `evaluate.py` code updated to assume missing explanations have rank of 10^9

Data split	Optimised TF-IDF	Iterated TF-IDF	BERT Re-ranking
Train	0.4525	0.4827	0.6677
Dev	0.4581	0.4966	0.5089
Test	0.4274	0.4576	0.4771
Time	0.02	46.97	92.96

Table 2: MAP scoring of new methods. The timings are in seconds for the whole dev-set, and the BERT Re-ranking f gure includes the initial Iterated TF-IDF step.

Analysis : By explanation length and type



Explanation role	Optimised TF-IDF	Iterated TF-IDF	BERT Re-ranking
GROUNDING	0.1373	0.1401	0.0880
LEX-GLUE	0.0655	0.0733	0.0830
CENTRAL	0.4597	0.5033	0.5579
BACKGROUND	0.0302	0.0285	0.0349
NEG	0.0026	0.0025	0.0022
ROLE	0.0401	0.0391	0.0439

Discussion

Original Baseline :

- Python version produces short output
- Evaluation requires 100% output
- TF-IDF method is un-optimised

Unlearned Methods :

- Use provided lemmatisation
- Systematically optimise TF-IDF
- Method 1 could be new baseline

BERT trained to rank :

- Gold explanations 'smooth' ranking
- Use pretrained LM to bridge gap and learn explanation strategy

Future directions :

- Still don't have solid grounding for Graph-based methods
- Formulate objective function to rate explanation sets

Source code available:

- http://RedDragon.ai/research

Key References

- "TextGraphs 2019 Shared Task on Multi-Hop Inference for Explanation Regeneration" - Jansen and Ustalov (2019)
- "Multi-hop inference for sentence-level textgraphs: How challenging is meaningfully combining information for science question answering?" - Jansen (2018)
- "A robustly optimized BERT pretraining approach" - Liu et at. (2019)

Contact

martin@RedDragon.ai
+65 8585 1750
http://RedDragon.ai