NAACL 2021

TextGraphs 2021 Shared Task System Description

Multi-Hop Inference Explanation Regeneration by Matching Expert Ratings

Vivek Kalyan
hello@vivekkalyan.com

Sam Witteveen
sam@RedDragon.ai

Martin Andrews
martin@RedDragon.ai

RED DRAGON AI

Summary

Shared Task :

- Rank explanation sentences for elementary school science questions to match 'Expert relevancy ratings'

Data Used :

- WorldTree V2 Corpus
- 250K gold Expert relevancy ratings

Ideas :

- Hyper-opt. BM25 incremental ranking
- Expert relevancy regression target
- Ensemble consistent output format

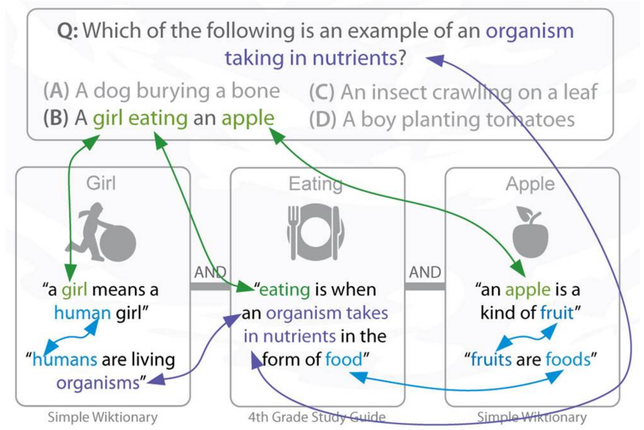
Results / Leaderboard Score :

- NDCG score : 0.7705 (ranked #2)

Key References

- "TextGraphs 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration" - Jansen and Ustalov (2020)
- "Red Dragon AI at TextGraphs 2020 shared task : LIT : LSTM-interleaved transformer for multi-hop explanation ranking" - Chia et al. (2020)
- "Learning to rank using gradient descent" - Burges et al. (2005)
- "BERT: Pre-training of deep bidirectional transformers for language understanding" - Devlin et al. (2019)

Task Setting



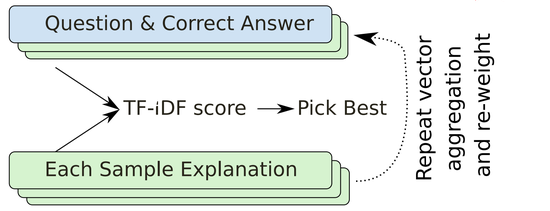
Shared Task this Year :

- Rank ~9000 explanation sentences to match 'expert relevancy ratings'

Pipeline 1 : Retrieval

"Iterated TF-IDF" improved (qv)

0.6583



- Initial retrieval tuned to return a 'manageable' list of initial guesses, with high recall

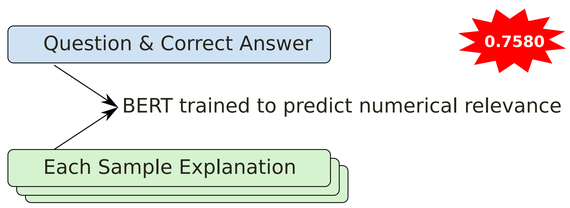
Retrieval Model	Oracle NDCG
TF-IDF	0.7547
I-BM25-base	0.8941
I-BM25	0.9378

Table 2: Oracle NDCG score on WorldTree V2 dataset

Pipeline 2 : LM Regression

Model Predicts Expert Relevance Score :

- Add a regression head to Transformer model and train on retrieved data



Language Model	Dev NDCG
DistilBERT	0.7353
BERT	0.7679
SciBERT	0.7541

Table 3: Language model comparison

Pipeline 3 : Ensembling

Results from different models Combined :

- Initially, sophisticated methods of ensembling were attempted
- Best results were from naïve score addition

0.7705

Model	Dev NDCG	Test NDCG
Baseline TF-IDF	0.5130	0.5010
I-BM25-base	0.6669	n/a
I-BM25	0.6785	0.6583
I-BM25 + BERT	0.7679	0.7580
I-BM25 + BERT ensemble	0.7801	0.7675
I-BM25 + BERT + SciBERT ensemble	0.7836	0.7705

Table 1: NDCG score comparison as evaluated locally and on the leaderboard

Discussion

New Task on Existing Dataset

- Previous tasks focussed on precision
- But this penalised the 'bigger picture'
- NDCG metric changes approach

Pipeline :

- I-BM25 method re-hyper-optimised
- Tried a number of BERT-like models
- Ensembling idea borrowed from 2020

Negative Results :

- Two-stage representation:
  - Relevant-or-not & Relevance Score
- Negative Sampling:
  - Address zero-relevance imbalance

Future directions :

- Again 'large language models' win!
- Simple metric changes can lead to very different modelling approaches
- Hope task returns to "reasoning" roots

Code & Contact

Source code is on GitHub, see:

- http://RedDragon.ai/research

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