Goal: How to evaluate truth value of natural language questions based on noisy textual evidence?

Will Ed Miliband stand down as the leader of the Labour party?

Labour leader Ed Miliband joins Instagram, is offered biscuit
Labour leader Ed Miliband is urged to go “all out for the win”

Ed Miliband is to blame for Johann Lamont quitting as the Scottish Labour Party leader
Ed Miliband resigns as Labour leader

Marrying Recognizing Textual Entailment and Information Retrieval

“Hypothesis Evaluation” Task: Binary natural language question or statement → multiple pieces of potential evidence → for each, determine the relevancy as well as degree and direction of entailment → produce a yes/no answer to the question.

Previous Work

Memory Networks: Trivial (bag-of-words) representations for retrieval + inference, with small vocabulary and simple clean sentences.

Answer Sentence Selection: Sophisticated models for retrieval, no inference.

SNLI: Large-scale RTE benchmark with sophisticated models for inference, but no retrieval.

HABCNN: Prior art CNNs on retrieval + inference, considers also sentence-level context.

Datasets (All Small)

Argus: Answering binary questions (on event occurrence) from news articles in a prediction market setting (retroactively).

AI2-8grade: High school science test (multiple choice) answered based on excerpts from CK12 textbooks. Kaggle task.


Arguments

Will Andre Iguodala win NBA Finals MVP in 2015?
Will Andre Iguodala lose the NBA Finals MVP award over Lebron James?
Will Donald Trump win the 2016 presidential election?

Will Donald Trump run for President in 2016?

Should Andre Iguodala have won the NBA Finals MVP award over LeBron James?

12:12 am ET Andre Iguodala was named NBA Finals MVP, not LeBron.

Should Andre Iguodala have won the MVP award over LeBron James?

Should Andre Iguodala have won the MVP award over LeBron James?

AI2-8grade

Sentence Pair Similarity: Question-answering task of finding the answers in a text.

(dataset-sts) package: Neural Network toolbox (Keras-based) that implements many popular models and tasks within this framework.

State-of-art e.g. on the Ubuntu Dialogue.

HABCNN has more information than we do, integrating surrounding sentences!

Live demo available (argus.ai2grade)

Open Source: https://github.com/brmson/dataset-sts


Neural Model

Sentence Embeddings

Sentence: Sequence of 50D GloVe word embeddings; 100 most frequent tokens kept trainable (operator words)

avg: MemNN-like bag-of-words

RNN: Bidirectional GRU

CNN: Multi-channel relu

Also tested RNN-CNN, attnS11

Pair score / (q, e): σ(w: q = e; q = e)!

Evidence Integration

Evidence Weighing: Use evidence embedding to predict attention-like relevance score in addition to the yes/no entailment score.

Evidence Averaging: No relevance model, only mean entailment.

BM25 Scoring: BM25 as another input for the entailment score.

Key trick: Pretraining sentence embeddings on the Ubuntu Dialogue dataset. All models are Siamese (same embedding for questions and evidence).

Evaluation

Alpha-8grade

Pretraining rocks! Deep learning can be applied to small datasets, opens up many practical applications.

We need better datasets with more unique events covered; practical performance does not match the 85% figure.

High variance of accuracy observed — multiple training runs are important!

Compositionality may be important - TreeReNN?

Acknowledgements

This research is supported by the CTU grant SGS16/084/OHK3/1T/13 and the Augur Project of the Forecast Foundation.

Department of Cybernetics, Czech Technical University, Prague