Joint Learning of Sentence Embeddings for Relevance and Entailment

Petr Baudiš, Silvestr Stanko and Jan Šedivý baudipet@fel.cvut.cz

Department of Cybernetics, Czech Technical University, Prague

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

Contribution: New neural network based schema for integrating and judging textual evidence.

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

Labour leader Ed Miliband joins Instagram, is offered biscuit	Ed Miliband is to blame for Johann Lamont quitting as the Scottish Labour Party leader
Labour leader Ed Miliband is urged to go "all out for the win"	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 occurence) 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.

MCTest: Multiple choice test on reading comprehension of short children stories. Standard benchmark.

Argus

Will Andre Iguodala win NBA Finals MVP in 2015?
Should Andre Iguodala have won the NBA Finals MVP award over LeBron James?
12.12am ET Andre Iguodala was named NBA Finals MVP, not LeBron.

Will Donald Trump run for President in 2016?
Donald Trump released "Immigration Reform that will make America Great Again" last weekend — his first, ...detailed position paper since announcing his campaign for the Republican nomination for president.
The Fix: A brief history of Donald Trump blaming everything on President Obama
DONALD TRUMP FOR PRESIDENT OF PLUTO!

AI2-8grade

***pedigree chart model* is used to show the pattern of traits that are passed from one generation to the next in a family?**
A pedigree is a chart which shows the inheritance of a trait over several generations.
Figure 51.14 In a pedigree, squares symbolize males, and circles represent females.

***energy pyramid model* is used to show the pattern of traits that are passed from one generation to the next in a family?**
Energy is passed up a food chain or web from lower to higher trophic levels.
Each step of the food chain in the energy pyramid is called a trophic level.

MCTest

It was Jessie Bear's birthday. She was having a party. She asked her two best friends to come to the party. ...
***one* / *two* / *six* / *four* friends came to Jessie's party.**
***Jessie Bear* / *no one* / *Lion* / *Tiger* was having a birthday.**

Sentence Pair Similarity

SPS model framework: Sentence embeddings (and IR baselines) for *comparing pairs of word sequences*.

Applicable Tasks: Paraphrasing, STS, RTE, answer sentence selection, ...

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.

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

Live demo available (argus.ailao.eu)

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

Neural Model

Sentence Embeddings

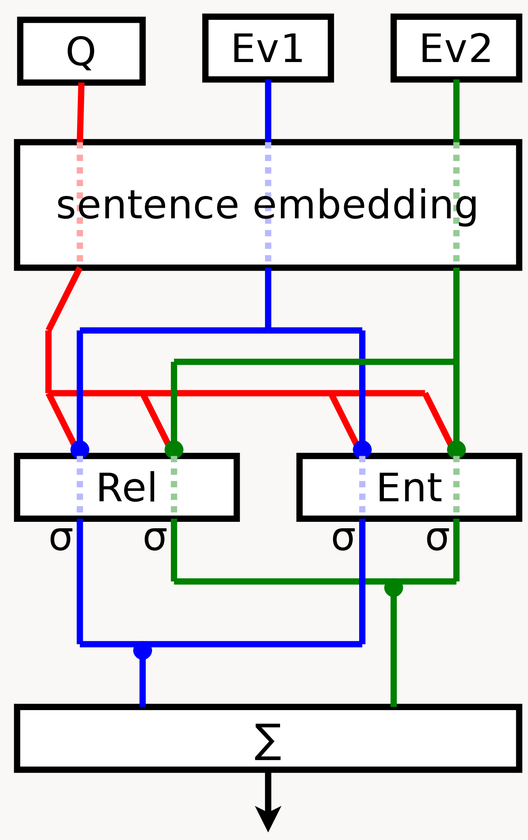
Sentence: Sequence of 50D GloVe word embeddings; 100 most frequent tokens kept trainable ("operator" words)

- **avg:** MemNN-like bag-of-words
- **RNN:** Bidirectional GRU
- **CNN:** Multi-channel relu
- Also tested RNN-CNN, *attn1511*

Pair score $f(\mathbf{q}, \mathbf{e})$: $\sigma(\mathbf{w}_f \cdot [\mathbf{q}+\mathbf{e}; \mathbf{q} \odot \mathbf{e}])$

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

Argus

Model	train	val	test
avg	0.872 ±0.009	0.816 ±0.008	0.744 ±0.020
RNN	0.906 ±0.013	0.875 ±0.005	0.823 ±0.008
CNN	0.896 ±0.018	0.857 ±0.006	0.822 ±0.007
Ubu. RNN	0.951 ±0.017	0.912 ±0.004	**0.852** ±0.008

AI2-8grade

Model	train	val	test
avg	0.505 ±0.024	0.442 ±0.022	0.401 ±0.016
RNN	0.712 ±0.053	0.381 ±0.016	0.361 ±0.012
CNN	0.676 ±0.056	0.442 ±0.012	0.384 ±0.011
Ubu. RNN	0.570 ±0.059	0.494 ±0.012	**0.441** ±0.011

MCTest

	MC-160			MC-500		
Model	one	multi	all	one	multi	all
hand-crafted	0.842	0.678	0.753	0.721	0.679	0.699
Attn. Reader	0.481	0.447	0.463	0.444	0.395	0.419
Neur. Reasoner	0.484	0.468	0.476	0.457	0.456	0.456
HABCNN-TE	0.633	**0.629**	**0.631**	0.542	**0.517**	0.529
avg	0.653 ±0.027	0.471 ±0.020	0.556 ±0.012	0.587 ±0.018	0.506 ±0.010	**0.542** ±0.011
RNN	0.583 ±0.033	0.490 ±0.018	0.533 ±0.020	0.539 ±0.016	0.456 ±0.013	0.494 ±0.012
CNN	0.655 ±0.020	0.511 ±0.012	0.578 ±0.014	0.571 ±0.013	0.483 ±0.012	0.522 ±0.009
Ubu. RNN	**0.736** ±0.033	0.503 ±0.016	**0.612** ±0.023	**0.641** ±0.017	0.452 ±0.017	**0.538** ±0.015
#1 Ubu. RNN	0.786	0.547	0.658	0.676	0.494	0.577

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

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.