Knowledge enhanced language models

Alexei Figueroa
MSc Data Science
Beuth University of Applied Sciences

Outline

- Motivation and problem statement
- Background and notation
- Data
- Methods and design choices
- Results
- Conclusions & Further work

Motivation and problem statement

Two men are in a room and the man with a blue shirt takes out a bench stone and with a little lubricant on the stone takes an knife and explains how to sharpen it. then he

a) uses a sharpener to smooth out the stone using the knife. (100.0%)
b) shows how to cut the bottom with the knife and place a tube on the inner and corner. (0.0%)
c) bends down and grabs the knife and remove the appliance. (0.0%)
d) stops sharpening the knife and takes out some pieces of paper to show how sharp the knife is as he cuts slivers of paper with the knife. (0.0%)

Figure 2.5: Context and endings from HellaSwag with their probabilities as evaluated with BERT, in red is BERT's answer and in bold the golden label. |Zellers et al., 2019b|

Figure 2.4: BERT validation accuracy when trained under several versions of SWAG and HellaSwag. BERT's performance on SWAG changes slightly, even with very aggressive alterations in the question structure, hinting to the presence of learned statistical features of the data instead of actual commonsense reasoning. |Zellers et al., 2019b|

Motivation and problem statement (continued)

Highly redundant models in SOTA -> repurpose parameters for more knowledge

Hypothesis

Given a **pretrained model**, such as, BERT-base-uncased, and a retraining with **multiple tasks** derived from commonsense **knowledge bases**, the retrained model will outperform a model of the same architecture without retraining when put through a commonsense **downstream task**.

Additionally, this retraining procedure will not affect significantly the capability of the model at general language tasks.

Background and notation (BERT)

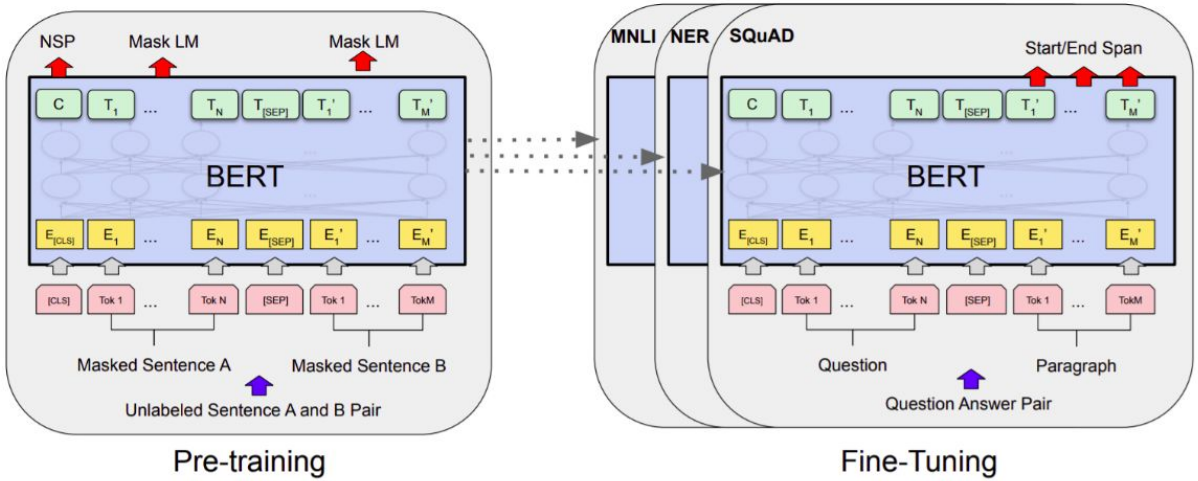
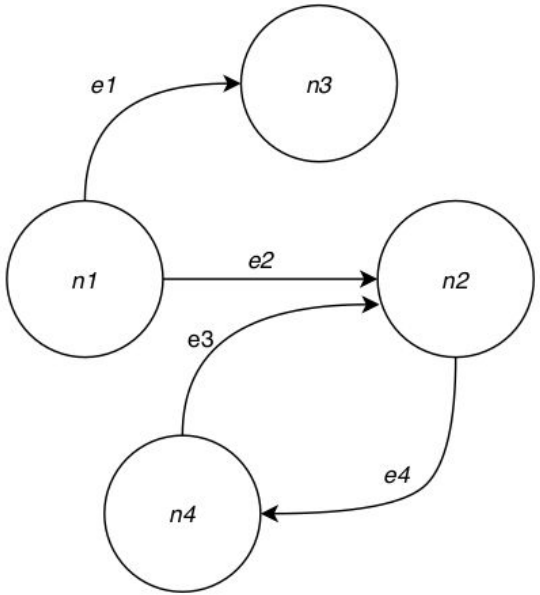


Figure 2.3: BERT, pre-training and finetuning settings differ only in the configuration of the output layers, `[CLS]` is a special token denoting the start of a sequence while `[SEP]` delimits two sequences, from [Devlin et al., 2018].

Background and notation (Knowledge Bases)

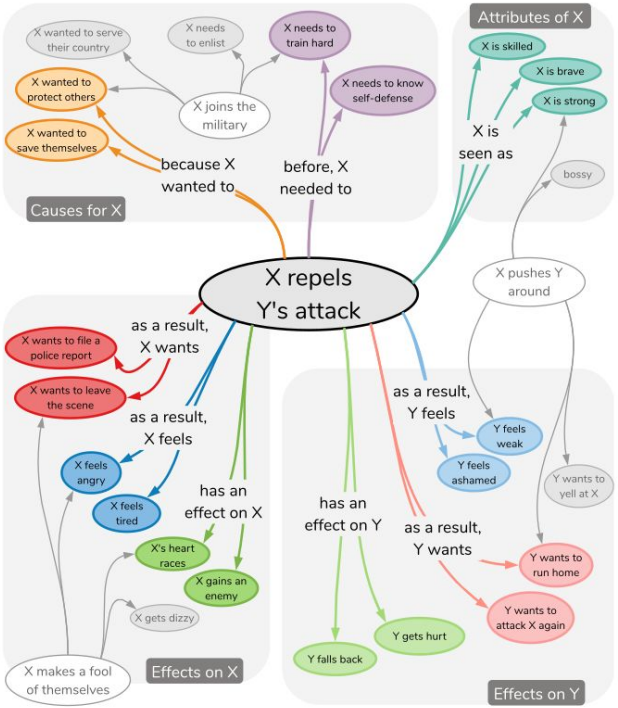


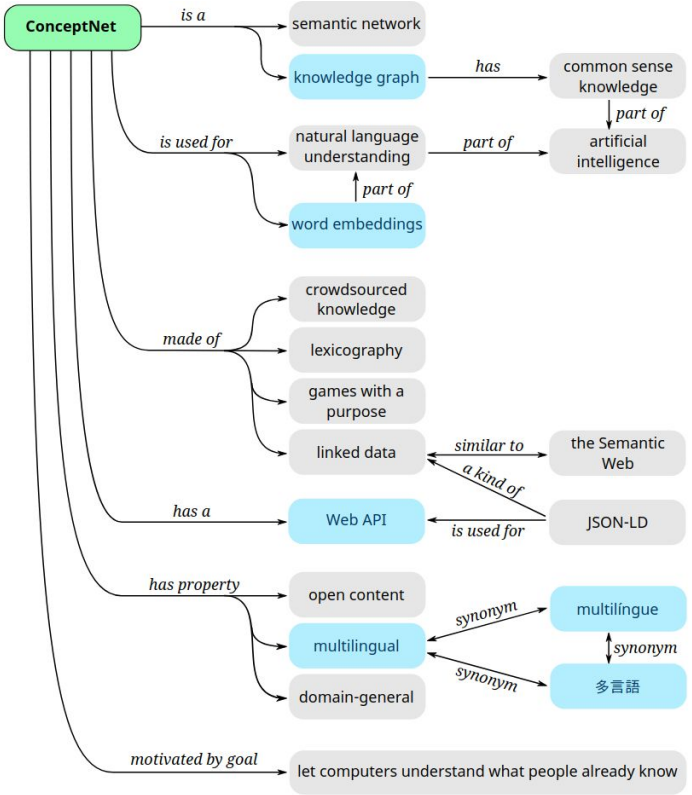
{ S, R, O }
{ n1, e1, n3 }
{ n1, e2, n2 }
{ n2, e4 ,n4 }
{ n4, e3, n2 }

Figure 3.2: $\{s, r, o\}$ triples in a sample directed graph, the nodes are qualified as s and o depending on the direction of the edge or relation r

(DATA) Knowledge Bases

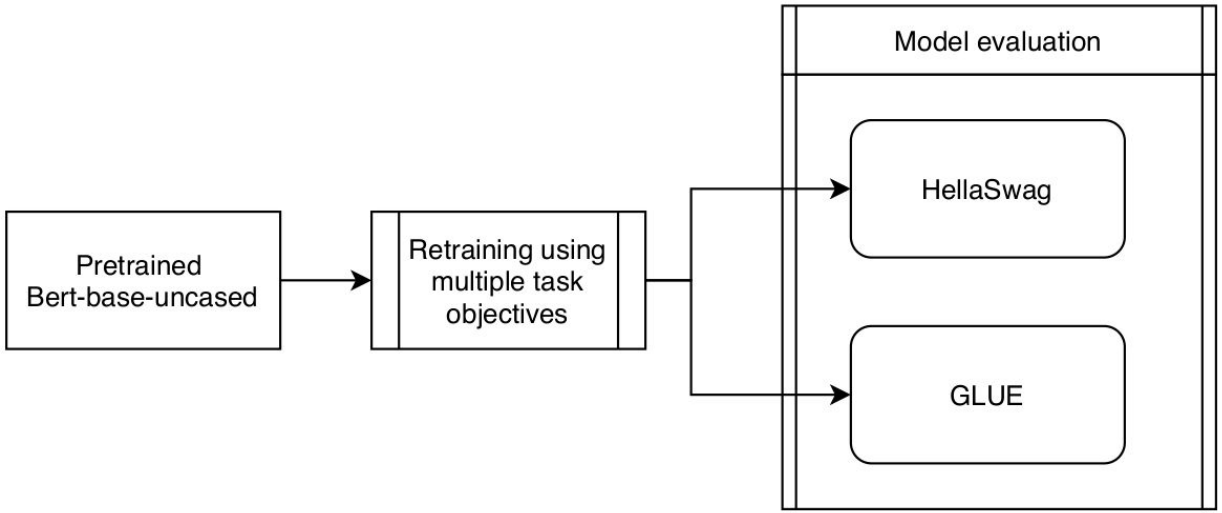
- **Atomic**: Inference (if-then) like relations
- **Conceptnet:** General knowledge base



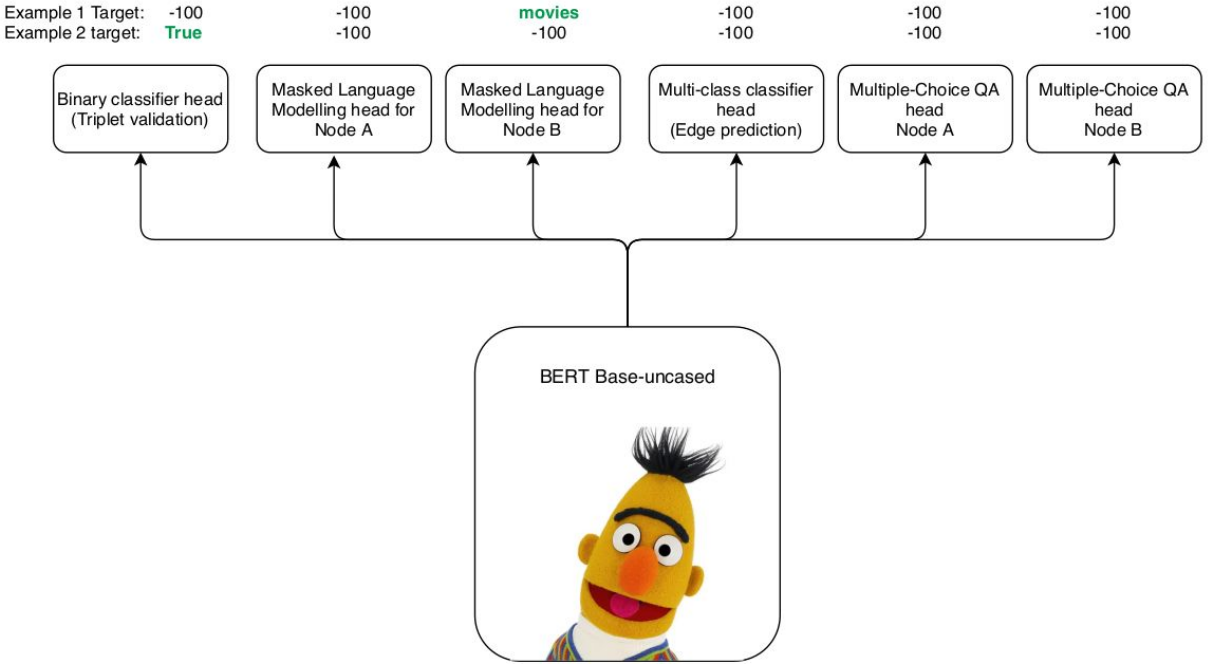


Methods and design choices (Overview)

Recipe for a proposed methodology to enhance a language model with knowledge
- a SOTA model, in this case BERT base.
- a domain specific downstream task, in this case HellaSwag.
- a domain specific knowledge base, in this case ATOMIC and CONCEPTNET.
- a multitask retraining routine.



Methods and design choices (Multitask model)



Methods and design choices (tasks)

Task	Task name	Evaluation metric
Triplet validation	IS_CORRECT	F1-Score
Edge prediction	MASK_EDGE	F1-Score
Masked language modelling of s	MASK_A	Perplexity
Masked language modelling of o	MASK_B	Perplexity
Multiple choice of s	MC_NODE_A	F1-Score
Multiple choice of o	MC_NODE_B	F1-Score

Table 3.1: Tasks of the retraining, their naming and the metrics used to control them

Implementation

CLUSTER

kubernetes

Results (HellaSwag)

Basline: Bert-base-uncased without retraining

	Acc	Loss	Datasets	Tasks	O Dropout	H Dropout	A Dropout
2020-06-18_135229	25.5%	1.386295	[CONCEPTNET]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.5	0.8	0.8
2020-06-10_093127	33.6%	1.475318	[CONCEPTNET]	[MASK_A, MASK_B, MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.25	0.1	0.1
2020-06-03_153757	34.1%	1.448410	[CONCEPTNET, ATOMIC]	[MASK_A, MASK_B, MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.25	0.1	0.1
2020-06-26_191749	34.7%	1.462585	[CONCEPTNET]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.9	0.1	0.1
2020-06-11_041946	34.8%	1.535658	[CONCEPTNET, ATOMIC]	[MASK_A, MASK_B, MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.8	0.1	0.1
2020-06-09_072234	35.0%	1.517587	[CONCEPTNET, ATOMIC]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.25	0.1	0.1
2020-06-16_203240	35.4%	1.531923	[CONCEPTNET, ATOMIC]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.9	0.1	0.1
2020-06-16_154727	36.0%	1.489099	[ATOMIC]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.9	0.1	0.1
2020-06-15_120244	36.2%	1.570379	[CONCEPTNET, ATOMIC]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.8	0.1	0.1
2020-06-22_080804	36.5%	1.544366	[CONCEPTNET]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.25	0.1	0.1
2020-06-23_144602	36.7%	1.519188	[CONCEPTNET, ATOMIC]	[MASK_EDGE, IS_CORRECT, MC_NODEA, MC_NODEB]	0.9	0.1	0.1
bert-base-uncased	**38.3%**	2.044656	-	-	-	-	-

Table 5.3: Baseline and retrained model results on HellaSwag, **O Dropout**: Dropout applied to the multiple task output heads **H Dropout**: Dropout applied to the fully connected layers in the embeddings, encoder, and pooler. **A Dropout**: Dropout ratio for the attention.

Results (GLUE)

task model_name	Mcorr CoLA	Acc MNLI		F1 MRPC	Acc QNLI
2020-06-15-120244-42	0.46	0.83	0.77	0.85	0.89
2020-06-16-154727-673	0.38	0.81	0.73	0.83	0.88
2020-06-16-203240-673	0.37	0.81	0.76	0.84	0.88
2020-06-18-135229-673	0.00	0.35	0.68	0.81	0.51
2020-06-22-080804-673	0.48	0.82	0.78	0.85	0.90
2020-06-23-144602-673	0.41	0.81	0.80	0.86	0.88
2020-06-26-191749-141	0.26	0.80	0.78	0.85	0.87
bert-base-uncased	**0.57**	**0.85**	**0.86**	**0.90**	**0.91**

Table 5.4: Baseline and retrained results for the GLUE benchmark, tasks CoLA, MNLI, MRPC and QNLI

Results (GLUE)

task Model Name	Acc QQP	F1 QQP	Acc RTE	SST-2	Pearson STS-B	Spearman STS-B	Acc WNLI
2020-06-15-120244	0.90	0.87	0.66	0.90	0.87	0.86	0.35
2020-06-16-154727-673	0.90	0.86	0.54	0.90	0.83	0.83	0.32
2020-06-16-203240-673	0.90	0.87	0.62	0.90	0.86	0.86	0.31
2020-06-18-135229-673	0.66	0.32	0.47	0.75	0.18	0.17	**0.56**
2020-06-22-080804-673	**0.91**	0.87	0.62	**0.92**	0.87	0.86	0.46
2020-06-23-144602-673	0.90	0.86	**0.65**	0.90	0.87	0.86	0.39
2020-06-26-191749-141	0.89	0.86	0.53	0.89	0.83	0.82	0.46
bert-base-uncased	**0.91**	**0.88**	**0.65**	**0.92**	**0.89**	**0.88**	0.46

Table 5.5: Baseline and retrained results for the GLUE benchmark, tasks QQP, RTE, SST-2, STS-B and WNLI

Error Analysis

Model name	Tr. Acc	Ev. Acc	Tr. Error	Ev. Error	Bias	Variance
2020-06-23_144602	71.6%	36.7%	28.4%	63.3%	24.0%	34.9%
bert-base-uncased	88.1%	38.3%	11.9%	61.7%	7.5%	49.7%

Table 5.6: Train and test error metrics for best retrained model and baseline.

	Baseline	
Retrained	Correct	Incorrect
Correct	2119	1567
Incorrect	1732	4624

Table 5.7: Comparison of the answers in the development set between the baseline raw pretrained BERT model and the retrained model 2020-06-23_144602

Model	Wins	Fails	Both right	Both wrong
Retrained	0.720902	1.772751	0.533695	2.102854
Baseline	2.502814	0.376710	0.255615	3.201260

Table 5.8: Comparison of the mean losses in the development set between the baseline raw pretrained BERT model and the retrained model 2020-06-23_144602. **Wins**: the retrained model chooses the correct answer and the baseline doesn't, **Fails**: the retrained model chooses the wrong answer while the baseline chooses the correct one. **Both right**: Both models choose correctly the answer and **Both Wrong**: Both models fail to choose a correct answer.

Error Analysis

Context : [CLS] the mother ins ##truct ##s them on how to brush their teeth while laughing. the boy helps his younger sister brush his teeth

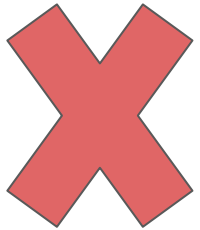
y	Retrained	Baseline	Ending
0	0.02	0.02	she shows how to hit the mom and then kiss his dad as well.
0	0.47	0.08	she brushes past the camera , looking better soon after.
0	0.12	0.02	she glow ##s from the center of the camera as a reaction.
1	0.39	0.88	she gets them some water to ga ##rg ##le in their mouths.

Conclusions

Hypothesis

Given a pretrained model, such as, BERT-base-uncased, and a retraining with multiple tasks derived from commonsense knowledge bases, the retrained model will outperform a model of the same architecture without retraining when put through a commonsense downstream task.

Additionally, this retraining procedure will not affect significantly the capability of the model at general language tasks.



Fairly close though

- Dropout
- HPO

Further work

- Improve on Masked Language Modelling task
- Selective parameter knowledge infusion
- Domain specific knowledge bases
- Hyper parameter optimization
- Additional models besides BERT
- Mozaic AI

?