

3/18 Paper Intro

+ Paper Report Questions

Pohan Chi

March 18, 2020

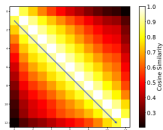
Outline

1. SBERT-WK
2. Mogrify LSTM
3. Retrospective Reader for Machine Reading Comprehension
4. Differentiable Reasoning over a Virtual Knowledge Base

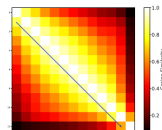
SBERT-WK: A Sentence Embedding Method By Dissecting BERT-based Word Models

Bin Wang, *Student Member, IEEE*, and C.-C. Jay Kuo, *Fellow, IEEE*

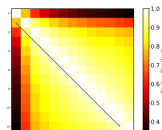
SBERT-WK - Motivation



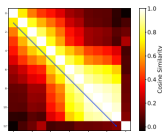
(a) BERT



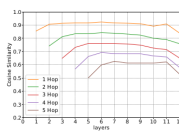
(b) SBERT



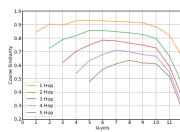
(c) RoBERTa



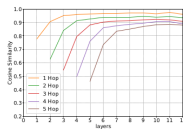
(d) XLNET



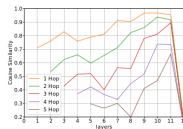
(e) BERT



(f) SBERT

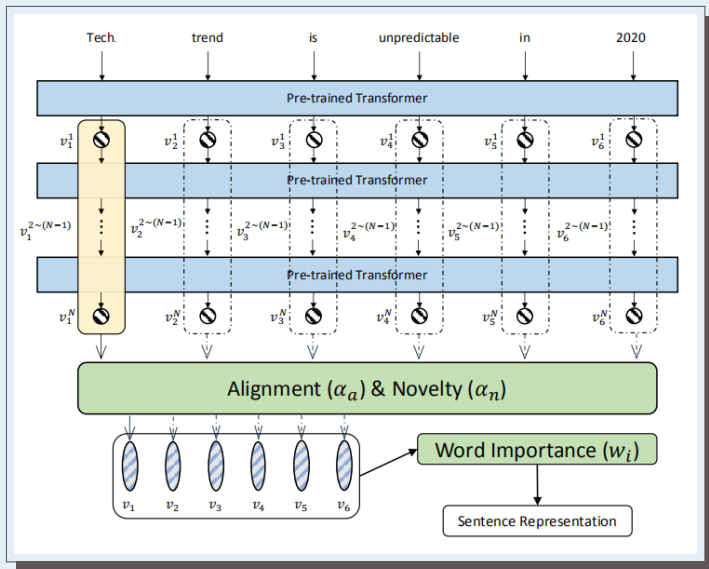


(g) RoBERTa



(h) XLNET

SBERT-WK - Model



SBERT-WK - Experiments

Dataset	Task	Sent A	Sent B	Label
STS12-STS16	STS	"Don't cook for him. He's a grown up."	"Don't worry about him. He's a grown up."	4.2
STS-B	STS	"Swimmers are racing in a lake."	"Women swimmers are diving in front of the starting platform."	1.6
SICK-R	STS	"Two people in snowsuits are lying in the snow and making snow angel."	"Two angels are making snow on the lying children"	2.5

Model	Dim	STS12	STS13	STS14	STS15	STS16	STSB	SICK-R	Avg.
<i>Non-Parameterized models</i>									
Avg. GloVe embeddings	300	52.3	50.5	55.2	56.7	54.9	65.8	80.0	59.34
Ave. FastText embedding	300	58.0	58.0	65.0	68.0	64.0	70.0	82.0	66.43
SIF	300	56.2	56.6	68.5	71.7	-	72.0	86.0	68.50
<i>p</i> -mean	3600	54.0	52.0	63.0	66.0	67.0	72.0	86.0	65.71
<i>Parameterized models</i>									
Skip-Thought	4800	41.0	29.8	40.0	46.0	52.0	75.0	86.0	52.83
InferSent-GloVe	4096	59.3	58.8	69.6	71.3	71.5	75.7	88.4	70.66
InferSent-FastText	4096	62.7	54.8	68.4	73.6	71.8	78.5	88.8	71.23
Universal Sentence Encoder	512	61.0	64.0	71.0	74.0	74.0	78.0	86.0	72.57
BERT [CLS]	768	27.5	22.5	25.6	32.1	42.7	52.1	70.0	38.93
Avg. BERT embedding	768	46.9	52.8	57.2	63.5	64.5	65.2	80.5	61.51
Sentence-BERT	768	64.6	67.5	73.2	74.3	70.1	74.1	84.2	72.57
Proposed SBERT-WK	768	70.2	68.1	75.5	76.9	74.5	80.0	87.4	76.09

SBERT-WK - Insight

MOGRIFIER LSTM

Gábor Melis[†], Tomáš Kočiský[†], Phil Blunsom^{†‡}

`{melisgl, tkocisky, pblunsom}@google.com`

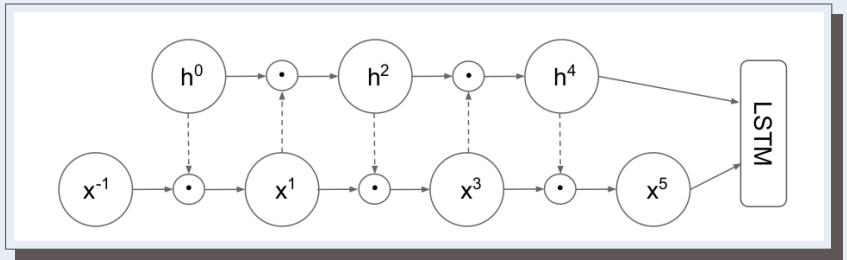
[†]DeepMind, London, UK

[‡]University of Oxford

Mogrify LSTM - Motivation

1. Improve Generalization ability of Language Model
2. Amplify salient and attenuate nuisance features in the input embeddings
3. Context-free representation is a bottleneck in LM
4. conditioning the input embedding on the recurrent state will improve performance.

Mogrify LSTM - Model



Mogrify LSTM - Experiments

			No Dyneval		Dyneval	
			Val.	Test	Val.	Test
PTB EN	FRAGE (d3, MoS15) (Gong et al. 2018)	22M	54.1	52.4	47.4	46.5
	AWD-LSTM (d3, MoS15) (Yang et al. 2017)	22M	56.5	54.4	48.3	47.7
	Transformer-XL (Dai et al. 2019)	24M	56.7	54.5		
	LSTM (d2)	24M	55.8	54.6	48.9	48.4
	Mogrifier (d2)	24M	52.1	51.0	45.1	45.0
	LSTM (d2, MC)	24M	55.5	54.1	48.6	48.4
	Mogrifier (d2, MC)	24M	51.4	50.1	44.9	44.8
WT2 EN	FRAGE (d3, MoS15) (Gong et al. 2018)	35M	60.3	58.0	40.8	39.1
	AWD-LSTM (d3, MoS15) (Yang et al. 2017)	35M	63.9	61.2	42.4	40.7
	LSTM (d2, MoS2)	35M	62.6	60.1	43.2	41.5
	Mogrifier (d2, MoS2)	35M	58.7	56.6	40.6	39.0
	LSTM (d2, MoS2, MC)	35M	61.9	59.4	43.2	41.4
	Mogrifier (d2, MoS2, MC)	35M	57.3	55.1	40.2	38.6

Mogrify LSTM - Insight

Retrospective Reader for Machine Reading Comprehension

Zhuosheng Zhang^{1,2,3}, Junjie Yang^{2,3,4}, Hai Zhao^{1,2,3,*},

¹Department of Computer Science and Engineering, Shanghai Jiao Tong University

²Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China

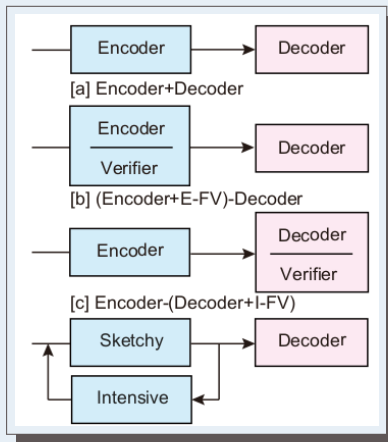
³MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

⁴SJTU-ParisTech Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, China
zhangzs@sjtu.edu.cn, jj-yang@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

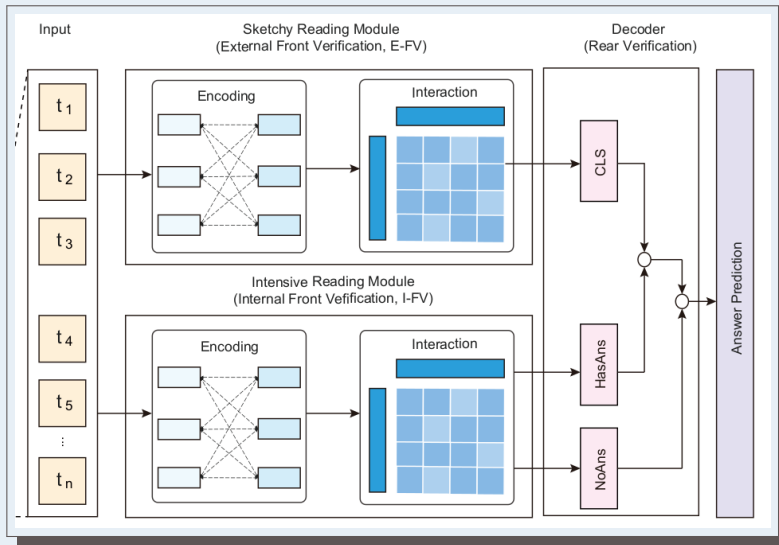
Retrospective Reader - Motivation

Roberta, XLNet, ALBERT - strong encoder !

How about focus on Decoder ?



Retrospective Reader - Model

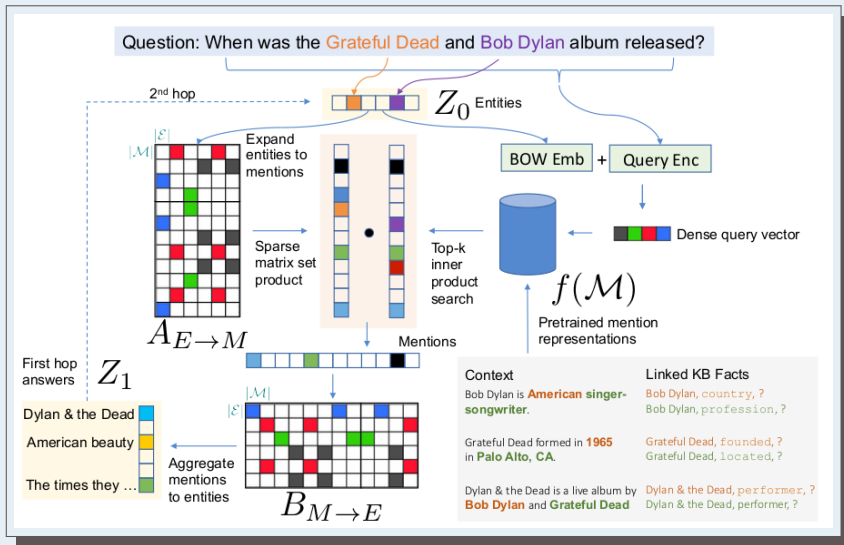


Retrospective Reader - Experiments

Model	Dev		Test	
	EM	F1	EM	F1
<i>Regular Track</i>				
Joint SAN	69.3	72.2	68.7	71.4
U-Net	70.3	74.0	69.2	72.6
RMR + ELMo + Verifier	72.3	74.8	71.7	74.2
<i>Top results on the leaderboard</i>				
Human	-	-	86.8	89.5
XLNet [Yang <i>et al.</i> , 2019]	86.1	88.8	86.4	89.1
RoBERTa [Liu <i>et al.</i> , 2019]	86.5	89.4	86.8	89.8
UPM [†]	-	-	87.2	89.9
XLNet + SG-Net Verifier++ [†]	-	-	87.2	90.1
ALBERT [Lan <i>et al.</i> , 2020]	87.4	90.2	88.1	90.9
ALBERT+ DA Verifier [†]	-	-	87.8	91.3
albert+verifier [†]	-	-	88.4	91.0
ALBERT (+TAV)	87.0	90.2	-	-
Retro-Reader over ALBERT	87.8	90.9	88.1	91.4

Retrospective Reader - Insight

Recap



Question 1

What is the meaning of color in Table A ?

What is the meaning of color in Table A ?

Ans:

The matrices in the figures are just illustrations and not accurate in any sense. But A is an $E \times M$ matrix, and $A[i, j] = 1$ implies that entity i co-occurs with mention j , as set by the threshold in equation 3. Since a single entity can co-occur with multiple mentions, and a single mention can co-occur with multiple entities, both your interpretations are correct.

Question 2

what is the initialization of z_0 ? (one hot or k-hot vector)

Answer 2

what is the initialization of z_0 ? (one hot or k-hot vector)

Ans:

We find a set of entities from the question and set them in z_0 .

In the case there is only 1 entity detected, it will be a 1-hot vector, otherwise it will be a k-hot vector.

Question 3

z will be a one-hot vector or a k -hot vector after first round ?

z will be a one-hot vector or a k-hot vector after first round ?

Ans: Probabilities in each dimension.