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

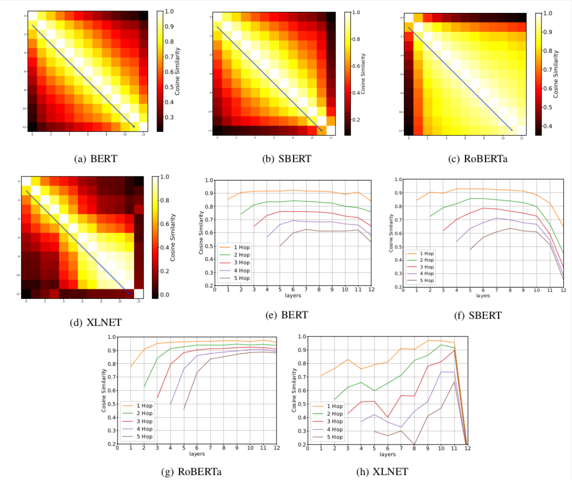
4. Differentiable Reasoning over a Virtual Knowledge Base

Title

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

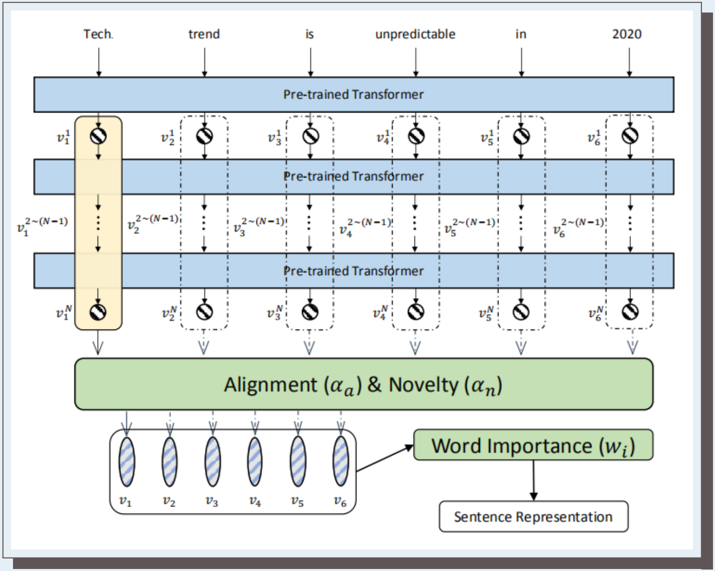


(a) BERT (b) SBERT (c) RoBERTa

(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.
Non-Parameterized models									
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
p-mean	3600	54.0	52.0	63.0	66.0	67.0	72.0	86.0	65.71
Parameterized models									
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

1. CLS token or averaging token embedding is good at Classification task but are not suitable for semantic textual similarity task.
2. Connection between Cosine similarity and NSP task.

Title

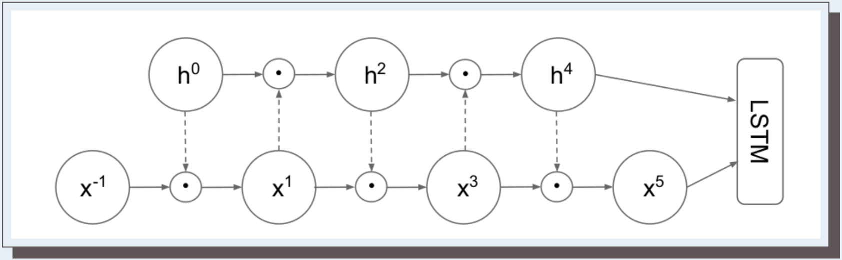
MOGRIFIER LSTM

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



$$x^i = 2\sigma(Q^i h_{prev}^{i-1}) \odot x^{i-2}$$ for odd i in [1,2,3,4..r]

$$h_{prev}^i = 2\sigma(R^i x^{i-1}) \odot h_{prev}^{i-2}$$ for even i in [1,2,3,4..r]

Mogrify LSTM - Experiments - (1)

			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 - Experiments - (2)

			No Dyneval		Dyneval	
			Val.	Test	Val.	Test
PTB EN	Trellis Networks (Bai et al. 2018)	13.4M		*1.159*		
	AWD-LSTM (d3) (Merity et al. 2017)	13.8M		1.175		
	LSTM (d2)	24M	1.163	1.143	1.116	1.103
	Mogrifier (d2)	24M	1.149	1.131	1.098	1.088
	LSTM (d2, MC)	24M	1.159	1.139	1.115	1.101
	Mogrifier (d2, MC)	24M	**1.137**	**1.120**	**1.094**	**1.083**
MWC EN	HCLM with Cache (Kawakami et al. 2017)	8M	*1.591*	*1.538*		
	LSTM (d1) (Kawakami et al. 2017)	8M	1.793	1.736		
	LSTM (d2)	24M	1.353	1.338	1.239	1.225
	Mogrifier (d2)	24M	1.319	1.305	1.202	1.188
	LSTM (d2, MC)	24M	1.346	1.332	1.238	NaN
	Mogrifier (d2, MC)	24M	**1.312**	**1.298**	**1.200**	**1.187**
MWC FI	HCLM with Cache (Kawakami et al. 2017)	8M	*1.754*	*1.711*		
	LSTM (d1) (Kawakami et al. 2017)	8M	1.943	1.913		
	LSTM (d2)	24M	1.382	1.367	1.249	1.237
	Mogrifier (d2)	24M	1.338	1.326	1.202	**1.191**
	LSTM (d2, MC)	24M	1.377	1.361	1.247	1.234
	Mogrifier (d2, MC)	24M	**1.327**	**1.313**	**1.198**	NaN
Enwik8 EN	Transformer-XL (d24) (Dai et al. 2019)	277M		**0.993**		**0.940**†
	Transformer-XL (d18) (Dai et al. 2019)	88M		1.03		
	LSTM (d4)	96M	1.145	1.155	1.041	1.020
	Mogrifier (d4)	96M	1.110	1.122	1.009	0.988
	LSTM (d4, MC)	96M	1.139	1.147		
	Mogrifier (d4, MC)	96M	1.104	1.116		
	Transformer-XL (d12) (Dai et al. 2019)	41M		1.06		1.01‡
	AWD-LSTM (d3) (Merity et al. 2017)	47M		1.232		
	mLSTM (d1) (Krause et al. 2016)	46M		1.24		1.08
	LSTM (d4)	48M	1.182	1.195	1.073	1.051
	Mogrifier (d4)	48M	1.135	1.146	1.035	1.012
	LSTM (d4, MC)	48M	1.176	1.188		
	Mogrifier (d4, MC)	48M	1.130	1.140		

Mogrify LSTM - Insight

1. LSTM v.s Transformer
2. another way to constrain LSTM input
3. Hypothesis

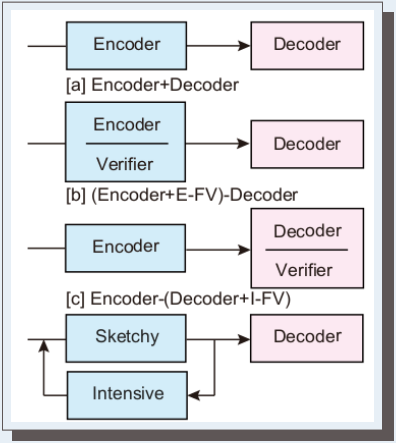
Title

Retrospective Reader for Machine Reading Comprehension

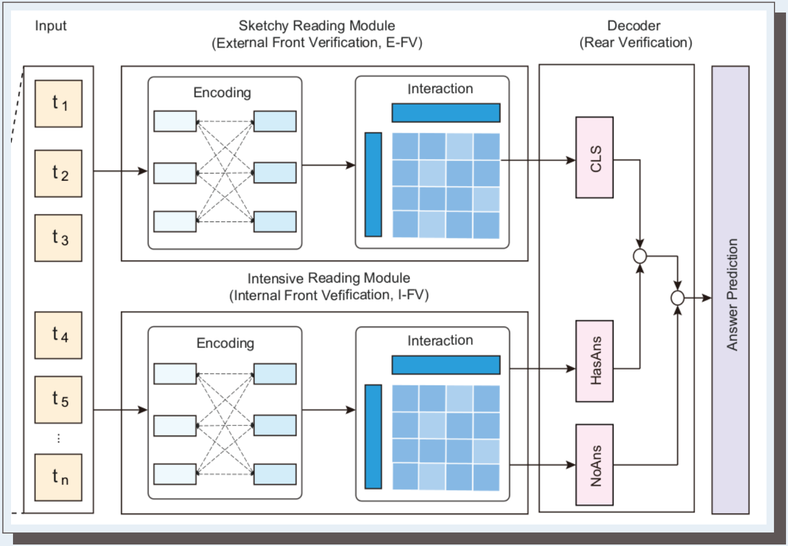
Zhuosheng Zhang[1,2,3], **Junjie Yang**[2,3,4], **Hai Zhao**[1,2,3,*],
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[4]SJTU-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 ?

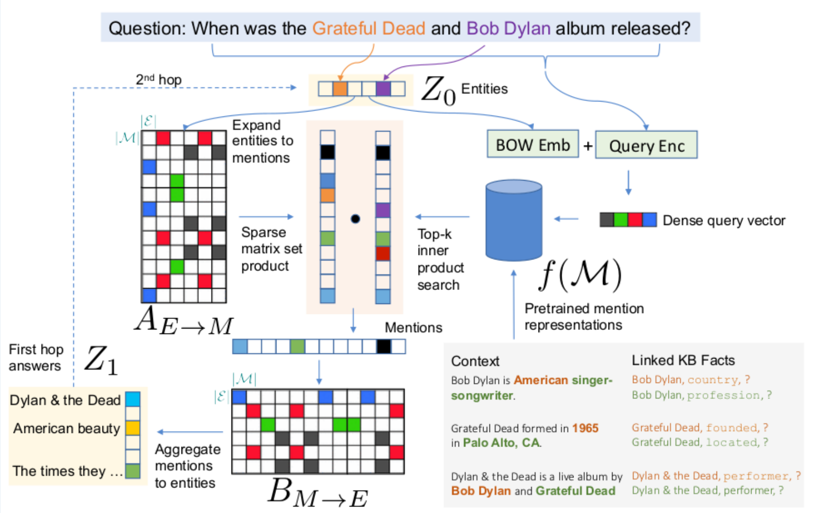


Retrospective Reader - Model



Retrospective Reader - Experiments

Model	Dev EM	Dev F1	Test EM	Test F1
Regular Track				
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
Top results on the leaderboard				
Human	-	-	86.8	89.5
XLNet [Yang *et al.*, 2019]	86.1	88.8	86.4	89.1
RoBERTa [Liu *et al.*, 2019]	86.5	89.4	86.8	89.8
UPM†	-	-	87.2	89.9
XLNet + SG-Net Verifier++†	-	-	87.2	90.1
ALBERT [Lan *et al.*, 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**



Question: When was the Grateful Dead and Bob Dylan album released?
2nd hop
Z_0 Entities
$|\mathcal{E}|$
$|\mathcal{M}|$
Expand entities to mentions
BOW Emb + Query Enc
Sparse matrix set product
Top-k inner product search
Dense query vector
$A_{E\to M}$
$f(\mathcal{M})$
Pretrained mention representations
Mentions
First hop answers
Z_1
Dylan & the Dead
American beauty
The times they …
Aggregate mentions to entities
$B_{M\to E}$
Context
Bob Dylan is American singer-songwriter.
Grateful Dead formed in 1965 in Palo Alto, CA.
Dylan & the Dead is a live album by Bob Dylan and Grateful Dead
Linked KB Facts
Bob Dylan, country, ?
Bob Dylan, profession, ?
Grateful Dead, founded, ?
Grateful Dead, located, ?
Dylan & the Dead, performer, ?
Dylan & the Dead, performer, ?

Question 1

What is the meaning of color in Table A ?

Answer 1

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 x 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 z0 ? (one hot or k-hot vector)

Answer 2

what is the initialization of z0 ? (one hot or k-hot vector)

Ans:
We find a set of entities from the question and set them in z0.
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 ?

Ans 3

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

Ans: Probabilities in each dimension.