Topic Compositional Neural Language Model

Wang et al 2018

AISTATS

Presenter Robert Thorstad

Problem: What Does *Magical* Mean?

1) Salaman Rushdie uses **magical** realism.

Context: fiction book

2) I watched Calvin Harris and it was **magical**

Context: music performance

3) Herbs have enormous **magical** power.

Context: book on wizardry

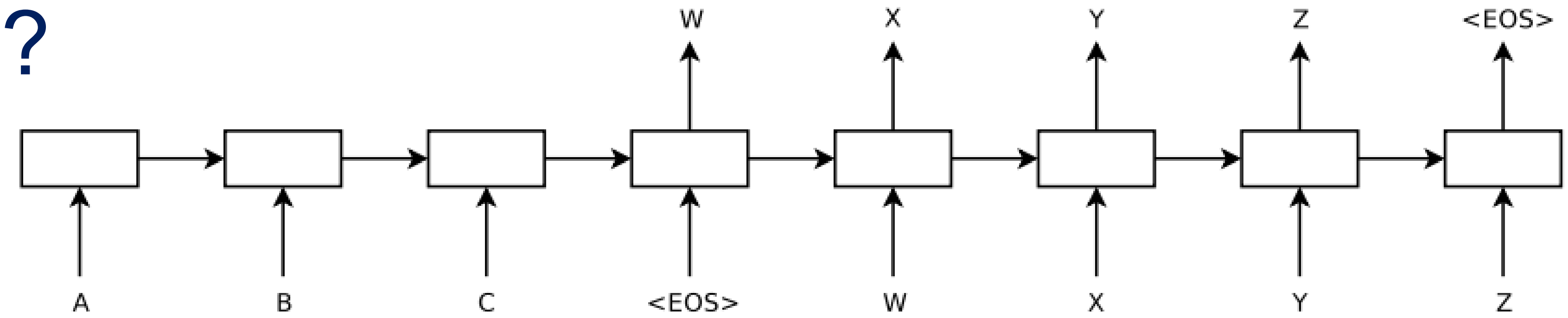
Need to know context!

(Ghosh et al, 2016)

Traditional Language Models Lack Long-Range Context

What came before? What is “it”?

?　　?　　?



It　was magical

(Sutskever, Vinyals, & Le, 2012)

Insight: Context Mostly Semantic

1) Salaman Rushdie uses **magical** realism.

Context: fiction book

2) I watched Calvin Harris and it was **magical**

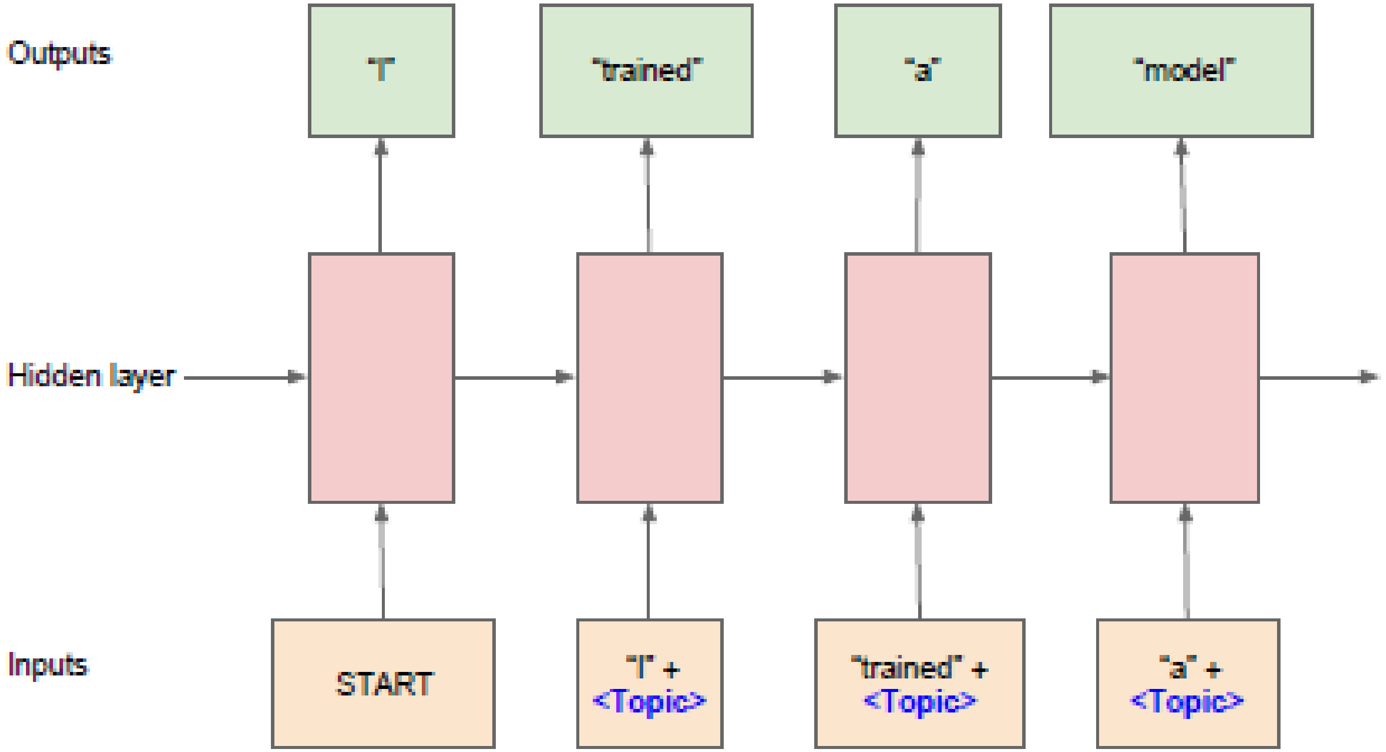
Context: music performance

3) Herbs have enormous **magical** power.

Context: book on wizardry

(Ghosh et al, 2016)

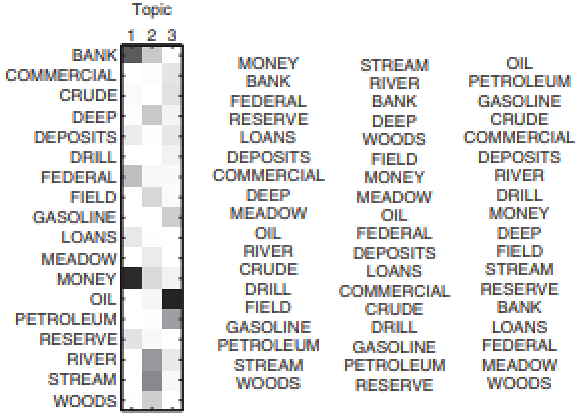
Idea: Add Context using Topic Model



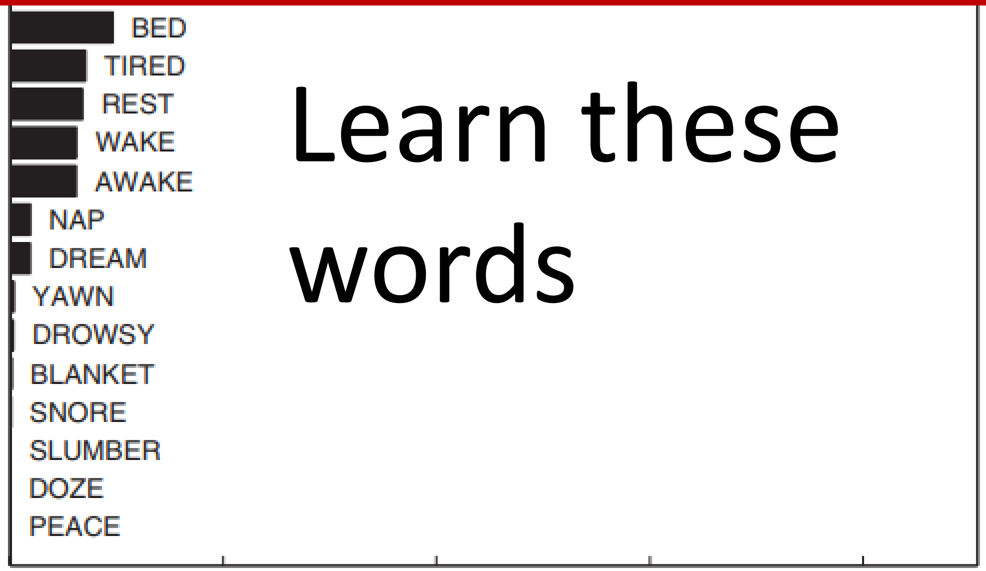
(Ghosh et al, 2016; similar idea by Mikolov et al, 2012)

Adding Topics Tells us Something About Psychology

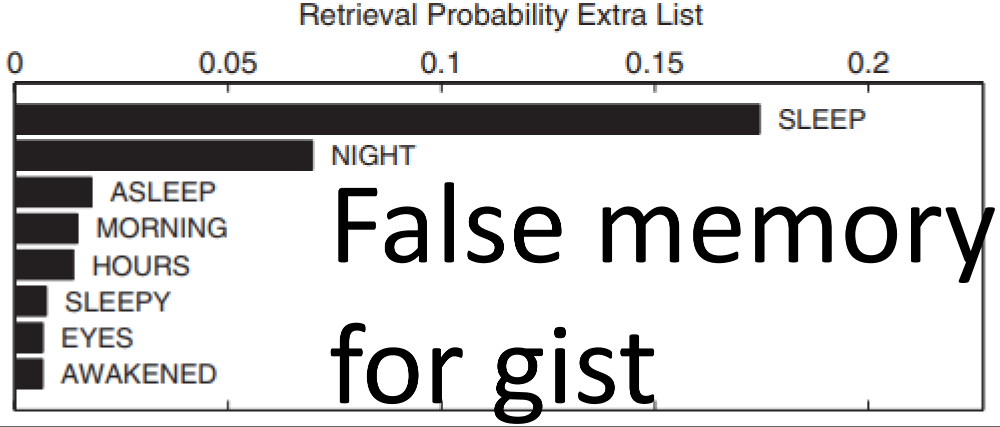
Topics as models of semantic knowledge base



(Griffiths, Steyvers, & Tenenbaum, 2007)



Learn these words



False memory for gist

Adding Topics Tells us Something About Psychology

People who can't remember past events:

"What did you do yesterday?"

Can't imagine future events:

"What are you going to do tomorrow?"

But can:

"What will be an important issue facing planet in next 10 years?"

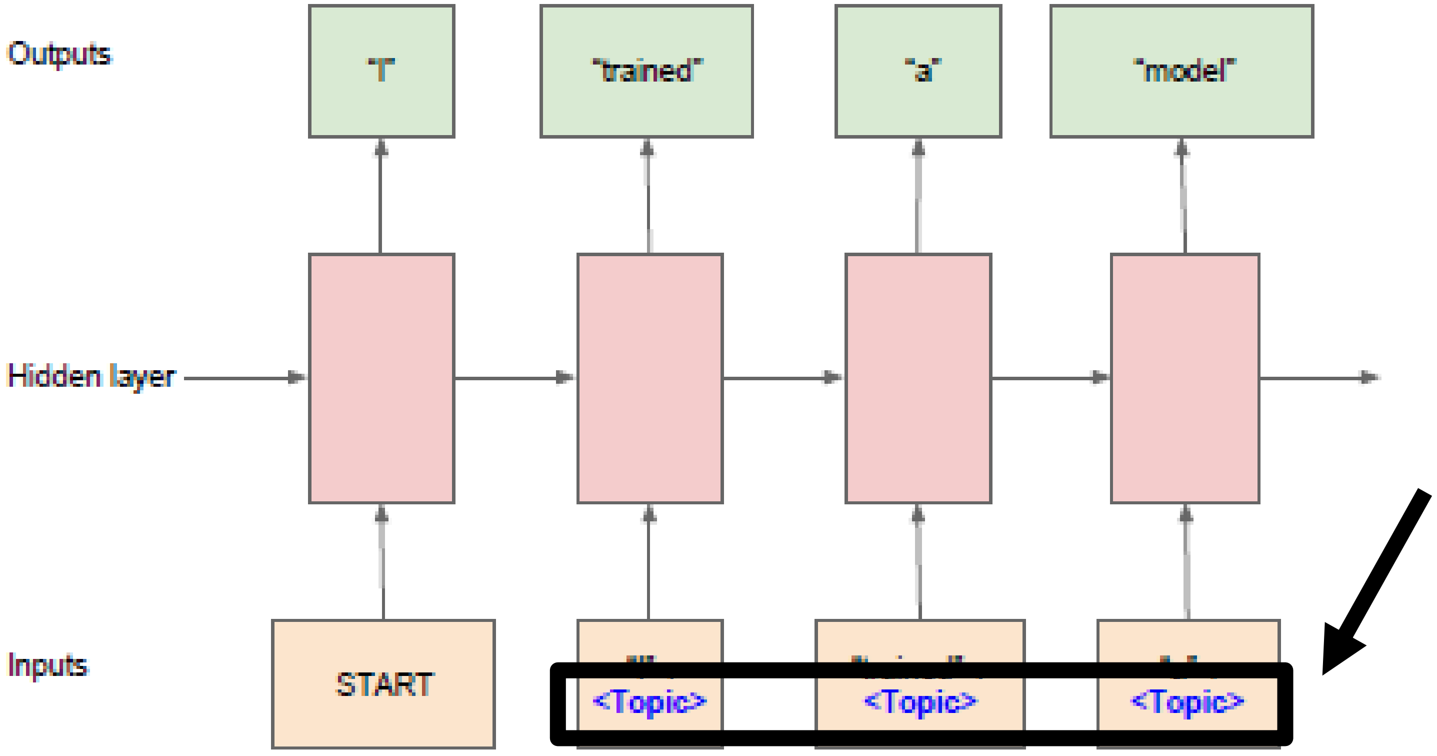
Wait for future rewards

"I believe a day should be planned ahead"

Using preserved semantic knowledge?

Klein et al 2002; Kwan et al 2013

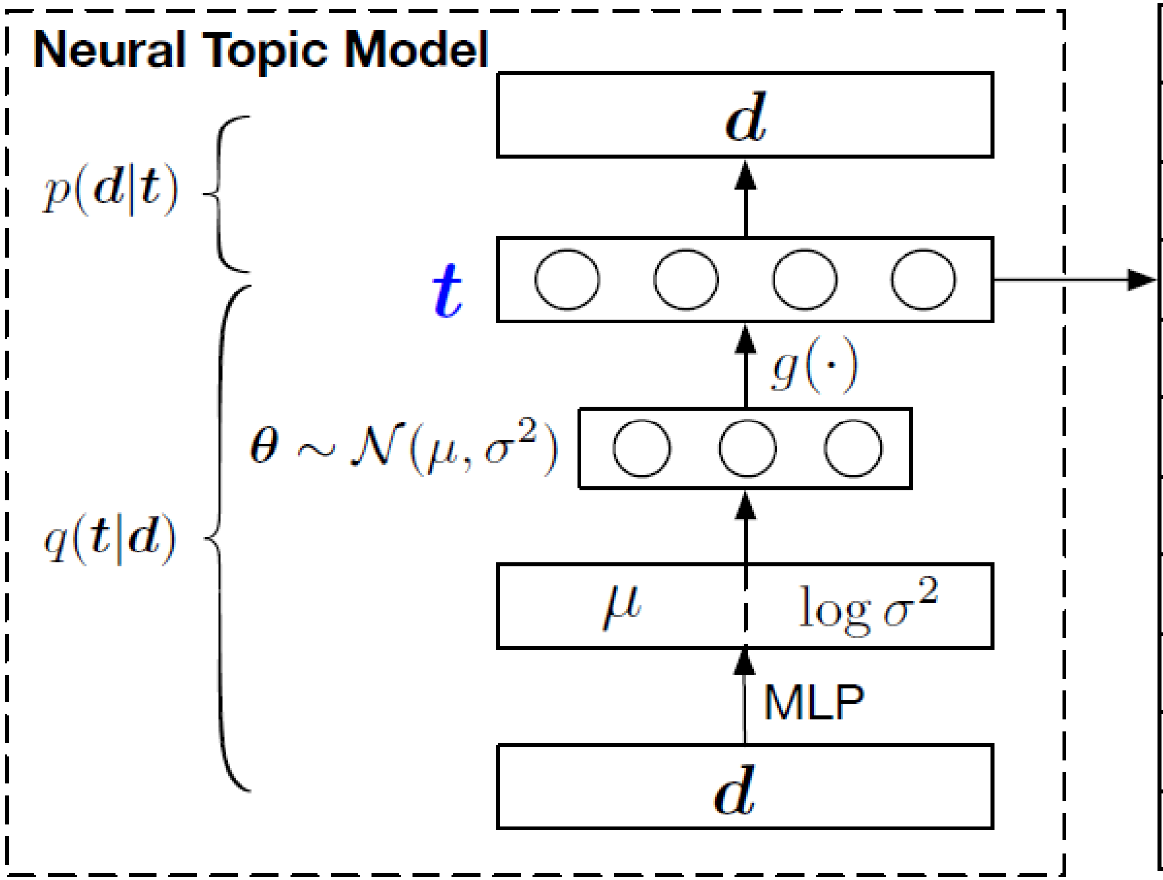
Really Want Topics Integrated in Model



(Ghosh et al, 2016)

Learned before training from a different corpus ☹

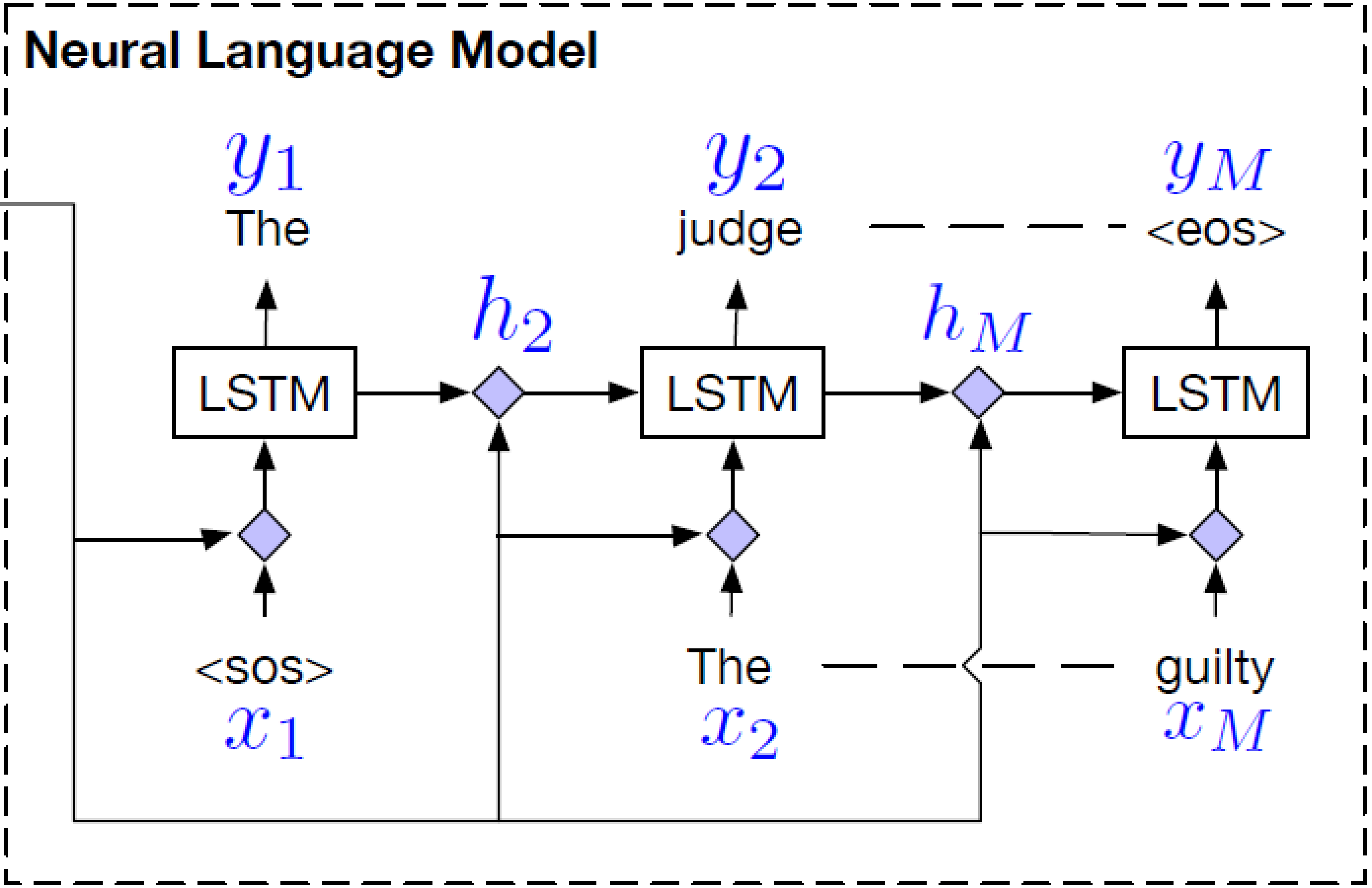
Proposal: Learn Topics using VAE...



Topic	Proportion
Law	0.36
Art	0.03
Market	0.10
Travel	0.07
Company	0.09
Politics	0.15
Sport	0.01
Education	0.11
Medical	0.02
Army	0.06

…Then Add Topics to Language Model

Topic	Proportion
Law	0.36
Art	0.03
Market	0.10
Travel	0.07
Company	0.09
Politics	0.15
Sport	0.01
Education	0.11
Medical	0.02
Army	0.06



Learns Qualitatively Good Topics

Dataset				
	army	**animal**	**medical**	**market**
APNEWS	afghanistan	animals	patients	zacks
	veterans	dogs	drug	cents
	soldiers	zoo	fda	earnings
	brigade	bear	disease	keywords
	infantry	wildlife	virus	share
	horror	**action**	**family**	**children**
IMDB	zombie	martial	rampling	kids
	slasher	kung	relationship	snoopy
	massacre	li	binoche	santa
	chainsaw	chan	marie	cartoon
	gore	fu	mother	parents
	environment	**education**	**politics**	**business**
BNC	pollution	courses	elections	corp
	emissions	training	economic	turnover
	nuclear	students	minister	unix
	waste	medau	political	net
	environmental	education	democratic	profits

Topics Generate Reasonable Sentences

Seed RNN with a topic or mixture (!) of topics:

horror	• the killer is a guy who is n't even a zombie .
action	• the action is a bit too much , but the action is n't very good .
family	• the film is also the story of a young woman whose <unk> and <unk> and very ye and palestine being equal , and the old man , a <unk> .
children	• i consider this movie to be a children 's film for kids .
war	• the documentary is a documentary about the war and the <unk> of the war .
horror+negative	• if this movie was indeed a horrible movie i think i will be better off the film .
sci-fi+children	• paul thinks him has to make up when the <unk> eugene discovers defeat in order and then finds his wife and boys .

Topics Competitive with Existing Models

Perplexity

Dataset	LSTM type	basic-LSTM*	LDA+LSTM*			LCLM*	Topic-RNN			TDLM*			TCNLM		
			50	100	150		50	100	150	50	100	150	50	100	150
APNEWS	small	64.13	57.05	55.52	54.83	54.18	56.77	54.54	54.12	53.00	52.75	52.65	52.75	52.63	**52.59**
	large	58.89	52.72	50.75	50.17	50.63	53.19	50.24	50.01	48.96	48.97	48.21	48.07	47.81	**47.74**
IMDB	small	72.14	69.58	69.64	69.62	67.78	68.74	67.83	66.45	63.67	63.45	63.82	63.98	62.64	**62.59**
	large	66.47	63.48	63.04	62.78	67.86	63.02	61.59	60.14	58.99	59.04	58.59	57.06	56.38	**56.12**
BNC	small	102.89	96.42	96.50	96.38	87.47	94.66	93.57	93.55	87.42	85.99	86.43	87.98	86.44	**86.21**
	large	94.23	88.42	87.77	87.28	80.68	85.90	84.62	84.12	82.62	81.83	80.58	80.29	80.14	**80.12**

Topics Competitive with Existing Models

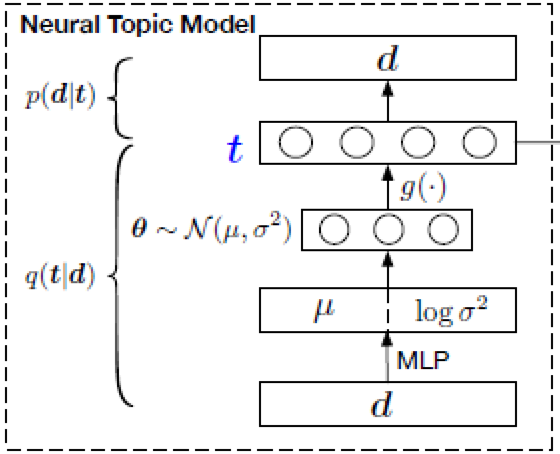
Coherence

# Topic	Model	Coherence APNEWS	Coherence IMDB	Coherence BNC
50	LDA*	0.125	0.084	0.106
	NTM*	0.075	0.064	0.081
	TDLM(s)*	0.149	0.104	0.102
	TDLM(l)*	0.130	0.088	0.095
	Topic-RNN(s)	0.134	0.103	0.102
	Topic-RNN(l)	0.127	0.096	0.100
	TCNLM(s)	**0.159**	**0.106**	**0.114**
	TCNLM(l)	0.152	0.100	0.101
100	LDA*	0.136	0.092	**0.119**
	NTM*	0.085	0.071	0.070
	TDLM(s)*	0.152	0.087	0.106
	TDLM(l)*	0.142	0.097	0.101
	Topic-RNN(s)	0.158	0.096	0.108
	Topic-RNN(l)	0.143	0.093	0.105
	TCNLM(s)	**0.160**	**0.101**	0.111
	TCNLM(l)	0.152	0.098	0.104
150	LDA*	0.134	0.094	**0.119**
	NTM*	0.078	0.075	0.072
	TDLM(s)*	0.147	0.085	0.100
	TDLM(l)*	0.145	0.091	0.104
	Topic-RNN(s)	0.146	0.089	0.102
	Topic-RNN(l)	0.137	0.092	0.097
	TCNLM(s)	0.153	**0.096**	0.107
	TCNLM(l)	**0.155**	0.093	0.102

Interestingly Absent:

Does adding topics make a better language model?

Need this… to do this?



Topic	Proportion
Law	0.36
Art	0.03
Market	0.10
Travel	0.07
Company	0.09
Politics	0.15
Sport	0.01
Education	0.11
Medical	0.02
Army	0.06