Explainable and Discourse Topic-aware Neural Language Understanding

Yatin Chaudhary [1 2] **Hinrich Schütze** [2] **Pankaj Gupta** [1]

A. Data Statistics and Evaluation

Table 1 shows data statistics of unlabeled and labeled datasets used to evaluate our proposed NCLM framework via Language Modeling (LM), Word Sense Disambiguation (WSD), Text Classification, Information Retrieval (IR) and Topic Modeling tasks. 20Newsgroups (20NS), Reuters (R21578) and AGnews are news-domain datasets which are labeled with 20, 90 and 4 classes respectively. Whereas, APNEWS, BNC are unlabeled news-domain datasets. However, IMDB movie reviews dataset (IMDB) is partially labeled i.e., 50K documents out of a total of 100K documents are labeled, with "positive" and "negative" sentiment labels in a single-label fashion. Therefore, we utilize all 100K documents for language modeling task and 50K labeled documents for information retrieval and text classification tasks on IMDB dataset. For Information Retrieval (IR) and Text Classification tasks, we utilize 20NS, AGnews and R21578 datasets along with 50K labeled documents from IMDB dataset. For Language Modeling (LM) and Topic Modeling (TM) tasks, we use unlabeled APNEWS, BNC, IMDB datasets and run experiments for a maximum of 100 epochs with early stopping criterion of 5 epochs.

B. Experimental setup for LM Evaluation

Table 2 shows the detailed hyperparameter settings for NTM and NLM components of our proposed NCLM framework for Language Modeling (LM) task. These settings are also utilized in Text Classification, Information Retrieval (IR) and Topic Modeling (TM) evaluations via best performing pretrained models on LM task. Based on the language modeling evaluation results from previous works for different number of topics i.e., $K \in \{50, 100, 150\}$, we fixed the number of topics $K = 150$ in the NTM component as this setting always performed best for related works.

[1]Corporate Technology, Machine Intelligence (MIC-DE), Siemens AG, Munich, Germany [2]CIS, University of Munich (LMU), Munich, Germany. Correspondence to: Yatin Chaudhary <yatin.chaudhary@drimco.net>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

C. Ablation study for α and $topN$

To select the best setting for hyperparameters α and $topN$ of our proposed NCLM framework, we perform an ablation study with $\alpha \in \{0.5, 0.1, 0.01\}$ and $topN \in \{10, 20, 40\}$ for each dataset as shown in Table 3 and select the settings with best language model perplexity scores for all of our experiments. We first find the best value of α by running experiments with LTA-NLM configuration and then freeze it to find the best value of $topN$ by running experiments with ETA-NLM configuration.

D. Time complexity of NCLM configurations

Based on the computational complexity formulation of the different configurations of our proposed NCLM framework described in section **3.6** in paper content, Table 4 shows the average run-time (in minutes) for one training epoch of our proposed models run on a NVIDIA Tesla K80 GPU with 12 GB memory. It is evident from Table 4 that *+SDT* configurations take much more time because of the computations of LTR/ETR vectors for all M words in sentence s.

E. Reproducibility: Code

Sections B and C describe the final hyperparameter settings we used in our evaluation experiments. To run the experiments and reproduce the scores reported in paper content, our implementation of NCLM framework is available at https://github.com/YatinChaudhary/NCLM. Due to the size of model parameters and datasets beyond upload limit, we have only provided code. Additional information such as raw/pre-processed datasets can be obtained, as detailed in the "README.md" file.

F. Qualitative topics and Text generation

For a qualitative evaluation of topic modeling component, Table 5 shows top 5 words of 10 selected topics extracted via NTM component for APNEWS, IMDB and BNC datasets. We further investigate the text generation capability of our proposed models by generating sentences conditioned on a particular topic signal as shown in Table 6. For a given topic k, during computation of latent $\mathbf{h}_d$ and explainable $\mathbf{z}_d^{att}$ topic representations of document d, we only utilize topic

Table 1. Preprocessed dataset statistics. Here, **#Docs** → documents, **#Sents** → sentences, "K" → thousand, and (∗) indicates vocabulary overlap with the corresponding vocabulary of APNEWS dataset for Information Retrieval (IR) and Text Classification tasks.

Datasets	**Train**		**Dev**		**Test**		**Vocabulary**		**Num**
	#Docs	**#Sents**	**#Docs**	**#Sents**	**#Docs**	**#Sents**	**NLM**	**NTM**	**Classes**
APNEWS	50K	662K	2K	275K	2K	264K	34230	7990	-
IMDB	75K	923K	12.5K	153K	12.5K	151K	36008	8714	-
BNC	15K	791K	1K	44K	1K	52K	43702	9741	-
20NS	9.9K	-	1K	-	7.5K	-	16884	6920	20
R21578	7.3K	-	0.5K	-	3K	-	8650	4943	90
AGnews	118K	-	2K	-	7.6K	-	21230	7500	4

Table 2. Hyperparameter settings of NCLM framework used in the experimental setup for Language Modeling (LM) task. Here, (∗) indicates hyperparameter values taken from the experimental setup of related work as mentioned under **Experimental Setup** in subsection **4.1** in paper content.

	Hyperparameter	**Value/Description**
NTM	$f^{MLP}*$	*1-layer* feed-forward neural network with 256 hidden units and *sigmoid* non-linearity
	l_1, l_2*	linear projections
	K*	150
	$topN$	[10, 20, 40]
	Pretraining epochs	20
NLM	Dropout probability*	0.4
	Max sequence length*	30
	small-NLM*	1-layer LSTM-LM with 600 hidden units
	large-NLM*	2-layer LSTM-LM with 900 hidden units
	Pretraining epochs*	10
	α	[0.5, 0.1, 0.01]
	Minibatch size*	64
	Learning rate*	0.001

Table 3. Ablation study over different settings of hyperparameters α and $topN$ for language modeling (LM) for APNEWS, IMDB and BNC datasets. For each dataset, **Bold** values indicate best LM perplexity scores and corresponding α & $topN$ hyperparameter settings are finalized for extensive LM experiments.

		APNEWS	**IMDB**	**BNC**
α	0.5	56.46	68.38	99.92
	0.1	55.97	**68.21**	98.85
	0.01	**55.48**	68.48	**98.31**
$topN$	10	49.26	60.65	99.63
	20	**49.34**	**59.20**	96.79
	40	51.26	61.95	**95.62**

proportion/key terms of kth topic and suppress participation from all other topics. Then we use these representations and a starting token "<bos>" to generate sentences in a greedy token-by-token fashion.

Table 4. Run-time for one epoch of our proposed models on APNEWS, IMDB, BNC datasets for language modeling task.

Model	**Run-time** (in minutes)		
	APNEWS	**IMDB**	**BNC**
LTA-NLM	45 ± 3	55 ± 3	50 ± 3
ETA-NLM	45 ± 3	55 ± 3	50 ± 3
LETA-NLM	45 ± 3	55 ± 3	50 ± 3
LTA-NLM +SDT	660 ± 15	780 ± 15	720 ± 15
ETA-NLM +SDT	660 ± 15	780 ± 15	720 ± 15
LETA-NLM +SDT	660 ± 15	780 ± 15	720 ± 15

Table 5. Top 5 words of 10 selected topics extracted from APNEWs, IDMB, BNC datasets.

APNEWS	**ethics**	**legal**	**election**	**weather**	**music**	**fraud**	**jail**	**fire**	**festival**	**robbery**
	lawsuit	jurors	republicans	storm	album	fraud	inmates	flames	tourism	prison
	complaint	trial	romney	winds	songs	scheme	corrections	drowned	visitors	arrested
	misconduct	execution	democrats	storms	guitar	laundering	prison	engulfed	event	pleaded
	violations	jury	nominee	flooding	music	fraudulent	jail	firefighters	organizers	stolen
	allegations	verdict	candidates	inches	film	restitution	probation	rescuers	celebration	theft
IMDB	**fiction**	**disney**	**animation**	**comedy**	**acting**	**bollywood**	**horror**	**thriller**	**sci-fi**	**religion**
	spock	daffy	disney	jokes	performance	khanna	cannibal	thriller	sci-fi	muslims
	batman	cindrella	animated	unfunny	supporting	saif	chainsaw	streep	alien	religion
	gundam	alladin	cartoons	satire	superb	khan	leatherface	hitchcock	spaceship	muslim
	superman	looney	kids	sandler	streep	amitabh	slasher	twists	science	jews
	joker	bambi	anime	snl	delivers	kapoor	zombie	mystery	predator	christianity
BNC	**novel**	**health**	**pollution**	**taxation**	**art**	**family**	**sports**	**air-force**	**expression**	**business**
	murder	hospital	emissions	council	paintings	women	goal	aircraft	eyes	corp
	police	care	environmental	cent	artist	mothers	scored	squadron	stared	ibm
	book	health	recycling	million	painting	parents	players	pilot	smiled	turnover
	story	nurses	waste	tax	museum	marriage	season	crew	looked	profits
	detective	staff	pollution	rates	gallery	child	league	battle	shook	sales

Table 6. Examples of sentences generated via our NCLM framework under the influence of unique topic signals. Key terms explaining each topic signal are presented in Table 5.

	TOPIC	GENERATED SENTENCE
APNEWS	**ethics**	the contract says the company will review the contract agreement with the company 's chief executive officer .
	fraud	prosecutors pleaded guilty in federal court in bank fraud conspiracy case .
	fire	the fire was reported sunday night in the town of <unk> , about 20 miles northeast of los angeles .
	festival	organizers will host events saturday at rhode island state park .
	robbery	authorities say officers arrested 24-year-old jose <unk> in las vegas on charges of robbery and assault in mexico after authorities say he shot his girlfriend in mexico in march 2012 .
IMDB	**thriller**	overall , it 's a solid thriller with plenty of action and action sequences , especially with a solid cast , solid performances , solid action sequences .
	comedy	i mean , if you are trying to laugh at jokes , please avoid this crap .
	animation	the animation is also quite impressive , but it 's not a visual achievement , but it 's a visual feast that is often overlooked in its own right .
	acting	she plays a young woman with a strong chemistry with her character , and she plays a role with a strong performance .
	thriller	however , it does not seem to reveal anything more than the plot , which is also quite effective .
BNC	**pollution**	the <unk> is a major source of energy , and the energy supply is not a waste of energy .
	taxation	the chancellor 's tax cuts are not a major factor in the rise in interest rates .
	art	the museum of art , sotheby 's , 9 june , est. $ 150,000 – 180,000 ; $ 440,000 – 180,000 ; $ 440,000 – <unk>) .
	novel	you can use the word ' <unk> ' to make a <unk> , but you can not be a detective .
	air-force	the <unk> aircraft , which is now operational , is expected to be upgraded to <unk> <unk> .