Supplementary Material

AAAI Press
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

Datasets

We use seven different datasets: (1) `20NSshort`: We take documents from 20NewsGroups data, with document size less (in terms of number of words) than 20. (2) `TREC6`: a set of questions (3) `Reuters21578title`: a collection of new stories from `nltk.corpus`. We take titles of the documents. (4) `Subjectivity`: sentiment analysis data. (5) `Polarity`: a collection of positive and negative snippets acquired from Rotten Tomatoes (6) `TMNtitle`: Titles of the Tag My News (TMN) news dataset. (7) `AGnewstitle`: Titles of the AGnews dataset. (8) `Reuters8`: a collection of news stories, processed and released by (9) `Reuters21578`: a collection of new stories from `nltk.corpus`. (10) `20NewsGroups`: a collection of news stories from `nltk.corpus`. (11) RCV1V2 (Reuters): `www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm` (12) `Sixxx Requirement OBjects` (`SiROBs`): a collection of paragraphs extracted from industrial tender documents (our industrial corpus).

The SiROBs is our industrial corpus, extracted from industrial tender documents. The documents contain requirement specifications for an industrial project for example, *railway metro construction*. There are 22 types of requirements i.e. class labels (multi-class), where a requirement is a paragraph or collection of paragraphs within a document. We name the requirement as Requirement Objects (ROBs). Some of the requirement types are *project management*, *testing*, *legal*, *risk analysis*, *financial cost*, *technical requirement*, etc. We need to classify the requirements in the tender documents and assign each ROB to a relevant department(s). Therefore, we analyze such documents to automate decision making, tender comparison, similar tender as well as ROB retrieval and assigning ROBs to a relevant department(s) to optimize/expedite tender analysis. See some examples of ROBs from SiROBs corpus in Table 1.

Experimental Setup and Hyperparameters for IR task

We set the maximum number of training passes to 1000, topics to 200 and the learning rate to 0.001 with $tanh$ hidden activation. We performed stochastic gradient descent based on the contrastive divergence approximation during 1000 training passes, with different learning rates. For model selection, we used the validation set as the query set and used the average precision at 0.02 retrieved documents as the performance measure. Note that the labels are not used during training. The class labels are only used to check if the retrieved documents have the same class label as the query document.

To perform document retrieval, we use the same train/development/test split of documents discussed in data statistics (experimental section) for all the datasets during learning. For model selection, we use the development set as the query set and use the average precision at 0.02 retrieved documents as the performance measure. We train DocNADE and iDocNADE models with 200 topics and perform stochastic gradient descent for 2000 training passes with different learning rates. Note that the labels are not used during training. The class labels are only used to check if the retrieved documents have the same class label as the query document. See Table 3 for the hyperparameters in the document retrieval task.

Experimental Setup and Hyperparameters for doc2vec model

Doc2Vec

We used gensim (`https://github.com/RaRe-Technologies/gensim`) to train Doc2Vec models for 12 datasets. Models were trained with distributed bag of words, for 1000 iterations using a window size of 5 and a vector size of 500.

Classification task

We used the same split in training/development/test as for training the Doc2Vec models (also same split as in IR task) and trained a regularized logistic regression classifier on the inferred document vectors to predict class labels. In the case of multilabel datasets (`R21578`,`R21578title`, `RCV1V2`), we used a one-vs-all approach. Models were trained with a liblinear solver using L2 regularization and accuracy and macro-averaged F1 score were computed on the test set to quantify predictive power.

Label: *training*
Instructors shall have tertiary education and experience in the operation and maintenance of the equipment or sub-system of Plant. They shall be proficient in the use of the English language both written and oral. They shall be able to deliver instructions clearly and systematically. The curriculum vitae of the instructors shall be submitted for acceptance by the Engineer at least 8 weeks before the commencement of any training.
Label: *maintenance*
The Contractor shall provide experienced staff for 24 hours per Day, 7 Days per week, throughout the Year, for call out to carry out On-call Maintenance for the Signalling System.
Label: *cables*
Unless otherwise specified, this standard is applicable to all cables which include single and multi-core cables and wires, Local Area Network (LAN) cables and Fibre Optic (FO) cables.
Label: *installation*
The Contractor shall provide and permanently install the asset labels onto all equipment supplied under this Contract. The Contractor shall liaise and co-ordinate with the Engineer for the format and the content of the labels. The Contractor shall submit the final format and size of the labels as well as the installation layout of the labels on the respective equipment, to the Engineer for acceptance.
Label: *operations, interlocking*
It shall be possible to switch any station Interlocking capable of reversing the service into "Auto-Turnaround Operation". This facility once selected shall automatically route Trains into and out of these stations, independently of the ATS system. At stations where multiple platforms can be used to reverse the service it shall be possible to select one or both platforms for the service reversal.

Table 1: SiROBs data: Example Documents (Requirement Objects) with their types (label).

Hyperparameter	**Search Space**
learning rate	[<u>0.001</u>, 0.005, 0.01]
hidden units	[50, 200]
iterations	[2000]
activation function	sigmoid
scaling factor	[True, <u>False</u>]

Table 2: Hyperparameters in Generalization evaluation in the DocNADE and iDocNADE for 50 and 200 topics. The <u>underline</u> signifies the optimal setting.

Hyperparameters for PPL and IR

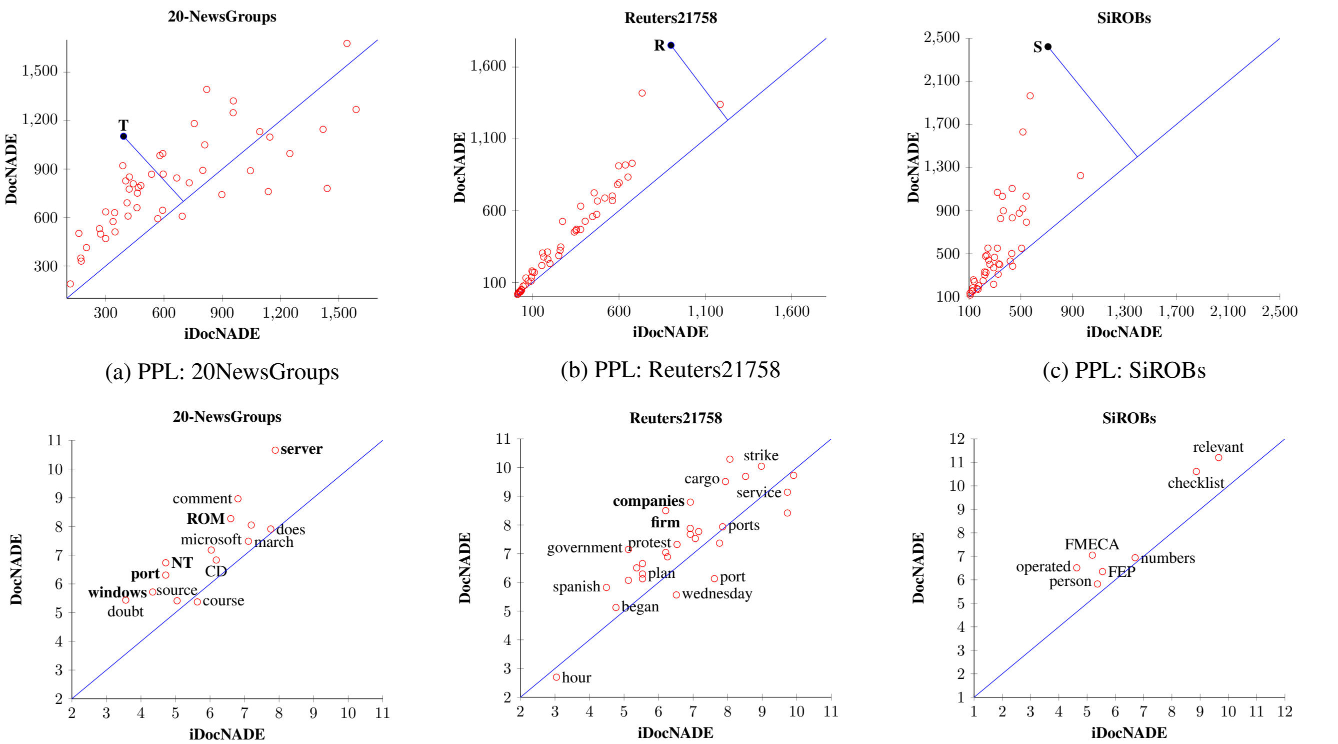
See Tables 2 and 3 for hyperparameters in generalization and retrieval tasks.

Hyperparameter	**Search Space**
retrieval fraction	[0.02]
learning rate	[<u>0.001</u>, 0.01]
hidden units	[200]
activation function	[<u>sigmoid</u>, tanh]
iterations	[<u>2000</u>, 3000]
scaling factor	[True, <u>False</u>]

Table 3: Hyperparameters in the Document Retrieval task. The <u>underline</u> signifies the optimal setting.

(from 20NewsGroups data set)
The CD-ROM and manuals for the March beta – there is no X windows server there.
Will there be? Of course. (Even) if Microsoft supplies one with NT , other vendors will no doubt port their's to NT.

Table 4: Raw text for the selected documents in generalization inspection from 20NewsGroups (20NS) dataset.



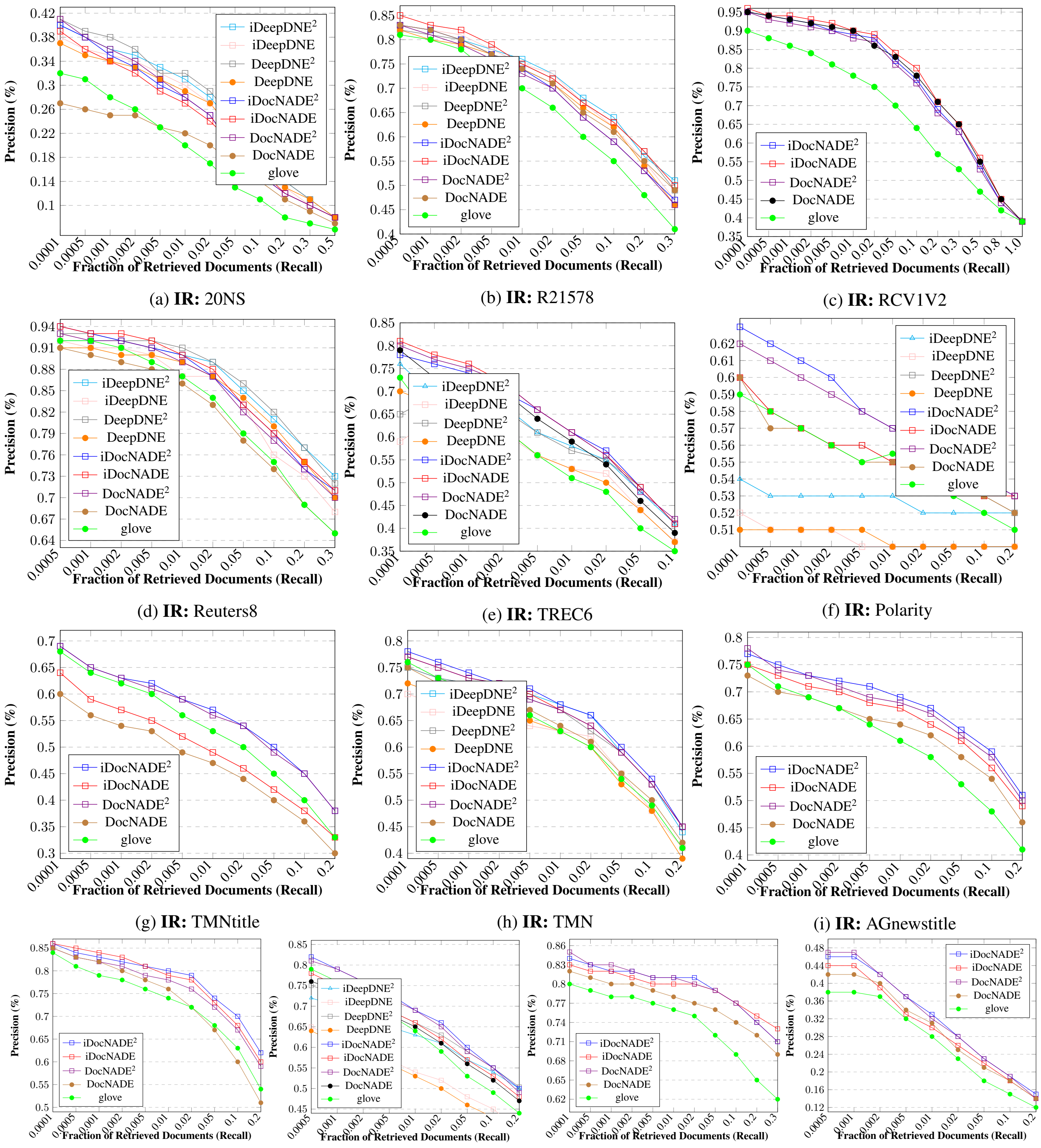
(a) PPL: 20NewsGroups (b) PPL: Reuters21758 (c) PPL: SiROBs

(d) NLL of 14 Words in **T** (e) NLL of 29 Words in **R** (f) NLL of 7 Words in **S**

Figure 1: (a, b, c): PPL (200 topics) by iDocNADE and DocNADE for each of the 50 held-out documents. The *filled circle* and symbols (**T**, **R** and **S** point to the document for which *PPL* differs by maximum, each for 20NewsGroups, Reuters21758 and SiROBs datsets, respectively. (d, e, f): NLL of each of the words in documents marked by **T**, **R** and **S**, respectively due to iDocNADE and DocNADE.

Data	**PPL**			
	DocNADE		**iDocNADE**	
	$T50$	$T200$	$T50$	$T200$
TREC6	42	42	39	<u>39</u>
Reuters8	178	172	162	<u>152</u>
Reuters21758	226	215	198	<u>179</u>
Polarity	310	311	294	<u>292</u>
20NewsGroups	864	830	836	<u>812</u>

Table 5: *PPL* and over 50 ($T50$) and 200 ($T200$) topics for DocNADE and iDocNADE using Tree-softmax. The *underline* indicates the best scores in PPL by iDocNADE.



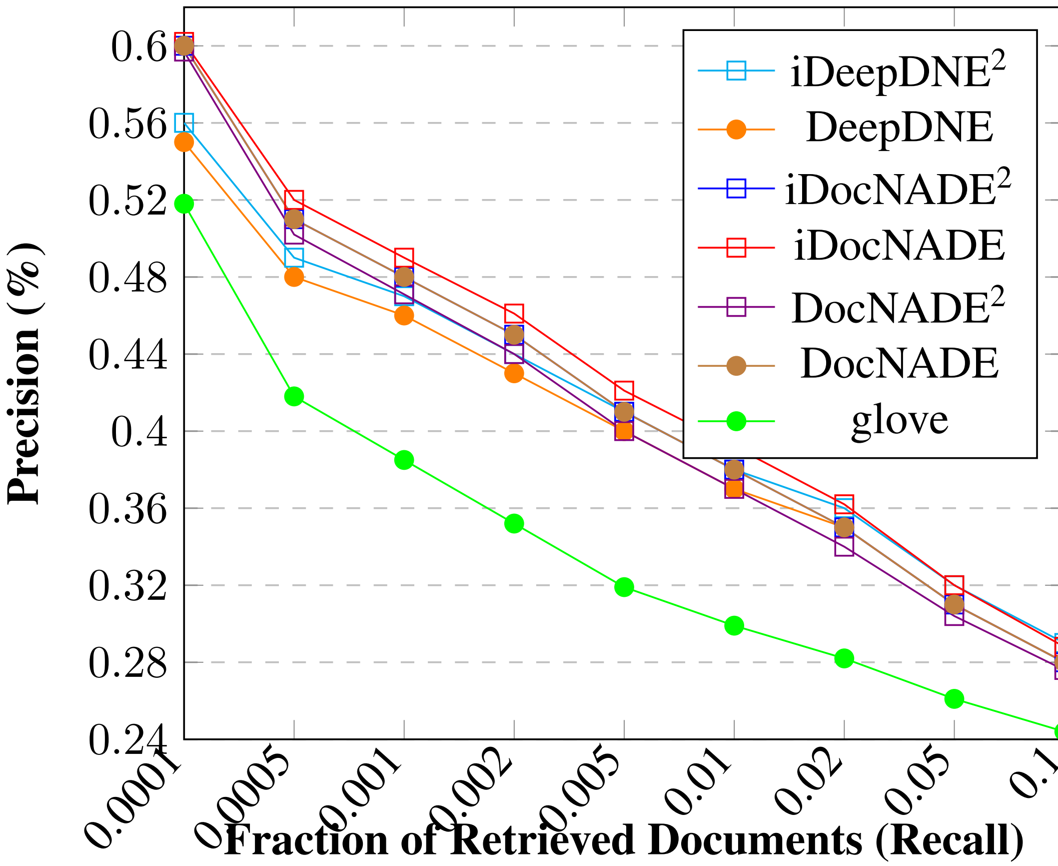
(a) **IR:** 20NS (b) **IR:** R21578 (c) **IR:** RCV1V2

(d) **IR:** Reuters8 (e) **IR:** TREC6 (f) **IR:** Polarity

(g) **IR:** TMNtitle (h) **IR:** TMN (i) **IR:** AGnewstitle

(j) **IR:** AGnews (k) **IR:** R21578title (l) **IR:** Subjectivity (m) **IR:** 20NSshort

Figure 2: Document retrieval performance (precision) at different retrieval fractions. Observe different y-axis scales.



(a) **IR:** SiROBs

Figure 3: Document retrieval performance (precision) at different retrieval fractions. Observe different y-axis scales.