hyperdoc2vec: Distributed Representations of Hypertext Documents

Jialong Han[♠], Yan Song[♠], Wayne Xin Zhao[♦], Shuming Shi[♠], Haisong Zhang[♠]

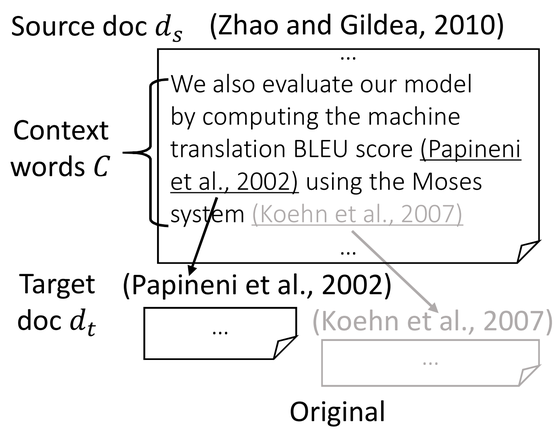
[♠] *Tencent AI Lab* [♦] *School of Information, Renmin University of China*

Motivation and Contributions

Embedding hyper-docs, *e.g.,* web pages and academic papers, may facilitate relevant tasks, *e.g.,* classification, recommendation, and retrieval.

- We propose four desired properties for hyper-document embedding models.
- We propose `hyperdoc2vec`, a general embedding approach for hyper-documents.
- We systematically validate the superiority of `h-d2v` in terms of the four properties.

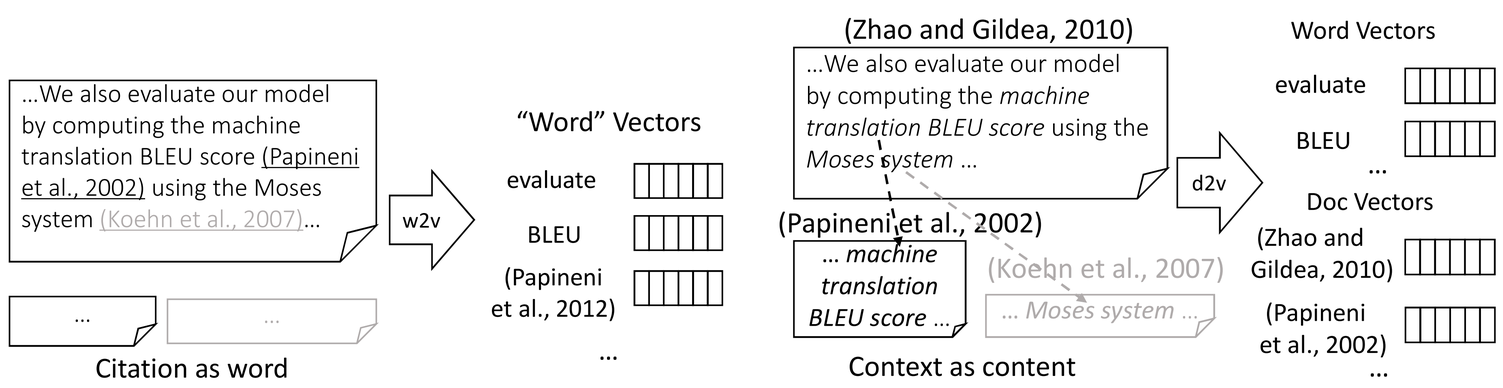
Notations and Problem Statement



Words and Document ID: $w \in W$ and $d \in D$.
Hyper-Document: a sequence on $W \cup D$ (of words and doc ids). Denote by H_d the hyper-document of document id d.
Hyperlink: a triple $\langle d_s, C, d_t \rangle$. $d_s, d_t \in D$ are ids of the *source* and *target* documents, respectively; $C \subseteq W$ are *context words*.
Embedding Vectors and Matrices: $\mathbf{w}, \mathbf{d}, \mathbf{W}$, and $\mathbf{D}$.

Input: a corpus of hyper-documents $\{H_d\}_{d \in D}$ with D and W.
Output: document and word embedding matrices $\mathbf{D} \in \mathbb{R}^{k \times |D|}$ and $\mathbf{W} \in \mathbb{R}^{k \times |W|}$.

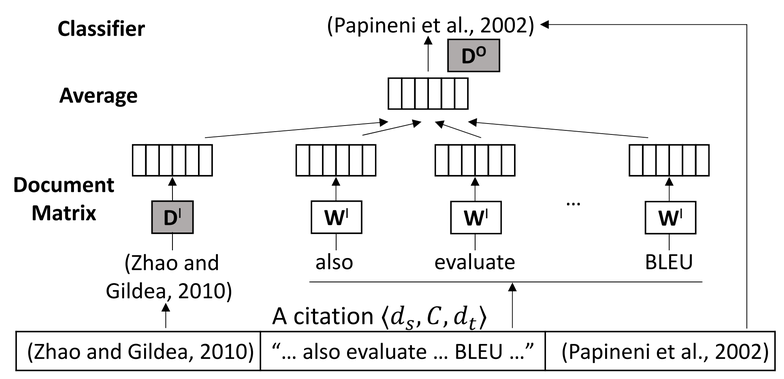
Adaptation of Existing Models



Desired Properties of Embedding Models

- **Content aware.** The document vector $\mathbf{d}$ of d should depend on words in H_d.
- **Context aware.** A hyperlink $\langle d_s, C, d_t \rangle$ should make $\mathbf{d}_t$ impacted by words in C.
- **Newcomer friendly.** Docs not referred to by any hyperlink should be embedded properly.
- **Context intent aware.** Words $w \in C$ indicates why d_s makes the reference (*e.g.,* "evaluate ... by" $\Rightarrow$ uses tools/algorithms in d_t). Their embeddings $\mathbf{w}$ should characterize such intents.

The `hyperdoc2vec` Model



Initialization: $\mathbf{D}^I, \mathbf{W}^I$, and $\mathbf{W}^O$, with `pv-dm`.

Optimization Objective:

$$\max_{\mathbf{D}^I, \mathbf{D}^O, \mathbf{W}^I} \quad \frac{1}{|\mathcal{C}|} \sum_{\langle d_s, C, d_t \rangle \in \mathcal{C}} \log P(d_t | d_s, C)$$

Probability of d_t Being Referred to:

$$P(d_t | d_s, C) = \frac{\exp(\mathbf{x}^\top \mathbf{d}_t^O)}{\sum_{d \in D} \exp(\mathbf{x}^\top \mathbf{d}^O)}$$

$$\text{where} \quad \mathbf{x} = \frac{1}{1 + |C|} \left(\mathbf{d}_s^I + \sum_{w \in C} \mathbf{w}^I \right)$$

Negative Sampling:

$$\log \sigma(\mathbf{x}^\top \mathbf{d}_t^O) + \sum_{i=1}^{n} \mathbb{E}_{d_i \sim P_N(d)} \log \sigma(-\mathbf{x}^\top \mathbf{d}_i^O)$$

Comparisons w.r.t. Desired Properties and Output

Desired Property	Impacts Task?		Addressed by Model?			
	Clf	CitRecom	w2v	d2v-nc	d2v-cac	h-d2v
Context aware	✓	✓	✓	×	✓	✓
Content aware	✓	✓	×	✓	✓	✓
Newcomer friendly	✓	✓	×	✓	✓	✓
Context intent aware	×	✓	×	×	×	✓

Model	Output			
	$\mathbf{D}^I$	$\mathbf{D}^O$	$\mathbf{W}^I$	$\mathbf{W}^O$
w2v	✓	✓	✓	✓
d2v (pv-dm)	✓	×	✓	✓
d2v (pv-dbow)	✓	×	×	✓
h-d2v	✓	✓	✓	✓

Experiments

Model	Original		+DeepWalk	
	Macro	Micro	Macro	Micro
DeepWalk	61.67	69.89	61.67	69.89
w2v (I)	10.83	41.84	31.06	50.93
w2v (I+O)	9.36	41.26	25.92	49.56
d2v-nc	70.62	77.86	70.64	78.06
d2v-cac	71.83	78.09	71.57	78.59
h-d2v (I)	68.81	76.33	**73.96**	**79.93**
h-d2v (I+O)	**72.89**	**78.99**	73.24	79.55

Model	NIPS				ACL Anthology				DBLP			
	Rec	MAP	MRR	nDCG	Rec	MAP	MRR	nDCG	Rec	MAP	MRR	nDCG
w2v (cbow, I4I)	5.06	1.29	1.29	2.07	12.28	5.35	5.35	6.96	3.01	1.00	1.00	1.44
w2v (cbow, I4O)	12.92	**6.97**	**6.97**	8.34	15.68	8.54	8.55	10.23	13.26	7.29	7.33	8.58
d2v-nc (pv-dbow, cosine)	14.04	3.39	3.39	5.82	21.09	9.65	9.67	12.29	7.66	3.25	3.25	4.23
d2v-cac (same as d2v-nc)	14.61	4.94	4.94	7.14	28.01	11.82	11.84	15.59	15.67	7.34	7.36	9.16
NPM (Huang *et al.*, 2015)	7.87	2.73	3.13	4.03	12.86	5.98	5.98	7.59	6.87	3.28	3.28	4.07
h-d2v (random init, I4O)	3.93	0.78	0.78	1.49	30.98	16.76	16.77	20.12	17.22	8.82	8.87	10.65
h-d2v (pv-dm init, I4O)	**15.73**	6.68	6.68	**8.80**	**31.93**	**17.33**	**17.34**	**20.76**	**21.32**	**10.83**	**10.88**	**13.14**

Data: NIPS (∼1.7k), ACL anthology (∼20k), and DBLP (∼650k). Citations parsed by ParsCit.

Tasks: Paper Classification (DBLP, 5,975 labels from Cora) and Citation Recommendation (all).

Observations (Paper Classification):

- `d2v-cac` > `d2v-nc` (✓ Context Aware).
- `w2v` has the worst performance (× Content Aware, × Newcomer Friendly).

Observations (Citation Recommendation):

- `d2v-cac` > `d2v-nc` (✓ Context Aware).
- NPM is sandwiched between `w2v`'s variants (× Context Aware, × Newcomer Friendly).
- `h-d2v` (`pv-dm` init) performs best. `pv-dm` init > random init (✓ Content Aware).

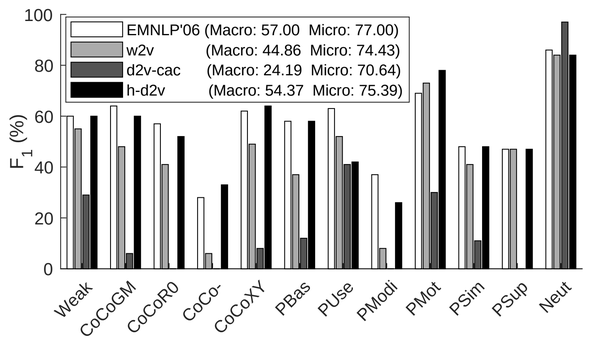
Remove Newcomers and Re-Classify:

- `w2v` benefits from intentionally screwing the data (× Newcomer Friendly).
- Even though the change, `w2v` still has the worst performance (× Content Aware).

Model	Original		+DeepWalk	
	Macro	Micro	Macro	Micro
DeepWalk	66.57	**76.56**	66.57	76.56
w2v (I)	19.77	47.32	59.80	72.90
w2v (I+O)	15.97	45.66	50.77	70.08
d2v-nc	61.54	73.73	69.37	78.22
d2v-cac	65.23	75.93	**70.43**	**78.75**
h-d2v (I)	58.59	69.79	66.99	75.63
h-d2v (I+O)	**66.64**	75.19	68.96	76.61

Classifying Citation Functions:

- 2,824 citation contexts with function labels, *e.g.,* PUse (using tools/algorithms), from Teufel *et al.* (2006). $\mathbf{w}^I$ as feature vector.
- `h-d2v` is the best non-feature-engineering approach (✓ Context Intent Aware).



Read our paper for more interesting studies and observations on Models × Properties!