P-SIF:Document Embeddings using Partition Averaging

Vivek Gupta[1,2], Ankit Saw[3], Pegah Nokhiz[1], Praneeth Netrapalli[2], Piyush Rai[4], Partha Talukdar[5]

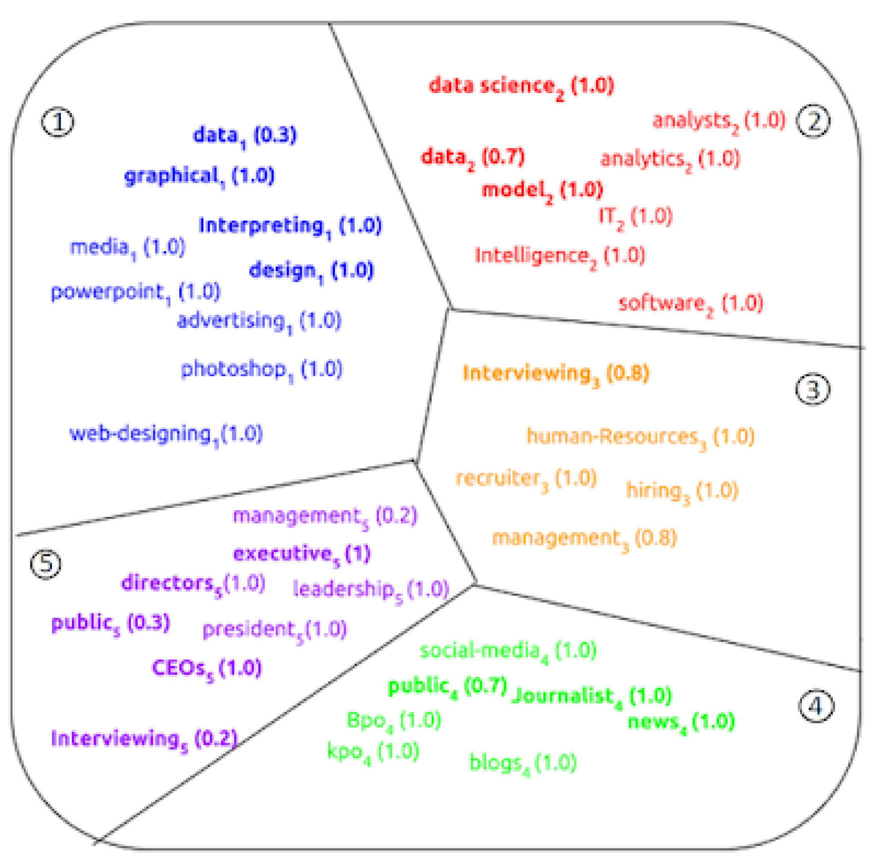
(1) University of Utah; (2) Microsoft Research Lab, India; (3) InfoEdge Ltd., India; (4) IIT Kanpur; (5) IISC, Bangalore

Distributional Semantics

- Each word (w) or sentence (s) is represented using a vector $\vec{v} \in \mathbb{R}^d$
- Semantically similar words or sentences occur closer in the vector space
- Various methods like word2vec (SGNS) and Doc2vec (PV- DBOW).

Averaging vs Partition Averaging

"Data journalists deliver data science news to the general public. They often take part in interpreting the data models. Also, they create graphical designs and interview the directors and CEOs."



- Direct Averaging to represent document

$$\vec{v}_{\text{data}_2} + \vec{v}_{\text{journalist}_4} + \vec{v}_{\text{news}_4} + \vec{v}_{\text{datascience}_1} + \vec{v}_{\text{public}_4} + \vec{v}_{\text{interpreting}_1} + \vec{v}_{\text{models}_2} + \vec{v}_{\text{graphical}_1} + \vec{v}_{\text{design}_1} + \vec{v}_{\text{directors}_5} + \vec{v}_{\text{CEO}_5} + \vec{v}_{\text{interviewing}_2}$$

- Partition Averaging to represent document

$$(\vec{v}_{interpreting} + \vec{v}_{graphical} + \vec{v}_{design}) \oplus (\vec{v}_{data} + \vec{v}_{datascience} + \vec{v}_{models}) \oplus (\vec{v}_{journalist} + \vec{v}_{news} + \vec{v}_{public}) \oplus (\vec{v}_{director} + \vec{v}_{CEO}) \oplus \vec{v}_{interviewing}$$

- Weighted Partition Averaging to represent document

$$(\vec{v}_{\text{interpreting}_1} + \vec{v}_{\text{graphical}_1} + \vec{v}_{\text{design}_1} + 0.3 * \vec{v}_{\text{data}_1}) \oplus (0.7 * \vec{v}_{\text{data}_2} + \vec{v}_{\text{datascience}_2} + \vec{v}_{\text{models}_2}) \oplus (\vec{v}_{\text{journalist}_4} + \vec{v}_{\text{news}_4} + 0.7 * \vec{v}_{\text{public}_4}) \oplus (\vec{v}_{\text{directors}_5} + 0.3 * \vec{v}_{\text{public}_5} + \vec{v}_{\text{CEO}_5} + 0.2 * \vec{v}_{\text{interviewing}_5}) \oplus 0.8 * \vec{v}_{\text{interviewing}_3}$$

Ways to Partition Vocabulary

Partition Type	Properties			
	Multi-Sense	Representation Sparsity	Non-Redundancy (Diversity)	Pre-Computation (Efficient)
Hard Clustering	✗	✓	✗	✗
Soft Clustering	✓	✗	✗	✓
Soft Clustering + Thresholding	✓	✓	✗	✗
Dictionary Learning	✓	✓	✓	✓

Ways to Represent Words

Embedding Type	Properties		
	Noise Robustness	Context Aware	Word Order-Syntax
SGNS	✗	✗	✗
Doc2VecC	✓	✗	✗
Multi-Sense + Doc2VecC	✓	✓	✗
BERT	✓	✓	✓

Kernels meet Embeddings

1. Simple Word Vector Averaging :

$$K^1(D_A, D_B) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle \vec{v}_{w_i^A} \cdot \vec{v}_{w_j^B} \rangle$$

2. TWE: Topical Word Embeddings :

$$K^2(D_A, D_B) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle \vec{v}_{w_i^A} \cdot \vec{v}_{w_j^B} \rangle + \langle \vec{tv}_{w_i^A} \cdot \vec{t}_{w_j^B} \rangle$$

3. P-SIF: Partition Word Vector Averaging :

$$K^3(D_A, D_B) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle \vec{v}_{w_i^A} \cdot \vec{v}_{w_j^B} \rangle \times \langle \vec{t}_{w_i^A} \cdot \vec{t}_{w_j^B} \rangle$$

4. Relaxed Word Mover Distance :

$$K^4(D_A, D_B) = \frac{1}{n} \sum_{i=1}^{n} \max_{j} \langle \vec{v}_{w_i^A} \cdot \vec{v}_{w_j^B} \rangle$$

Theoretical Justification of P-SIF

1. We provide theoretical justifications of P-SIF by showing connections with random walk-based latent variable models (Arora et al. 2016a; 2016b) and SIF embedding (Arora, Liang, and Ma 2017).

2. We relax one assumption in SIF to show that our P-SIF embedding is a strict generalization of the SIF embedding which is a special case with K = 1.

Text Similarity Task

	Document 1 (d_n^1)
Doc	A man is riding a motorcycle
SIF	$\vec{v}_{\text{man}_2} + \vec{v}_{\text{riding}_3} + \vec{v}_{\text{motorcycle}_4}$
P-SIF	$\vec{v}_{\text{zero}_1} \oplus \vec{v}_{\text{man}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{motorcycle}_4} \oplus \vec{v}_{\text{zero}_5}$

	Document 2 (d_n^2)
Doc	A woman is riding a horse
SIF	$\vec{v}_{\text{woman}_1} + \vec{v}_{\text{riding}_3} + \vec{v}_{\text{horse}_5}$
P-SIF	$\vec{v}_{\text{women}_1} \oplus \vec{v}_{\text{zero}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{zero}_4} \oplus \vec{v}_{\text{horse}_5}$

Similarity Scores

Ground Truth	weigh-Avg (SIF)	P-SIF
0.15	0.57	0.16

STS12	STS13	STS14	STS15	STS16
MSRpar	headline	deft forum	anwsers-forums	headlines
MSRvid	OnWN	deft news	answers-students	plagiarism
SMT-eur	FNWN	headline	belief	postediting
OnWN	SMT	images	headline	answer-answer
SMT-news		OnWN	images	question-question
		tweet news		

Model → Dataset ↓	PP -Proj	RNN	WME +PSL	Infer Sent	BERT (pr)	GRAN	Glove +WR	SIF +PSL	PSIF +PSL
STS12	60.0	58.4	62.8	61	53	62.5	56.2	59.5	**65.7**
STS13	56.8	56.7	56.3	56	**67**	63.4	56.6	61.8	64.0
STS14	71.3	70.9	68.0	68	62	**75.9**	68.5	73.5	74.8
STS15	74.8	75.6	64.2	71	73	**77.7**	71.7	76.3	77.3
STS16	-	64.9	-	**77**	67	-	72.4	72.5	73.7

Text Classification Task

- Multi-class text classification on 20NewsGroup

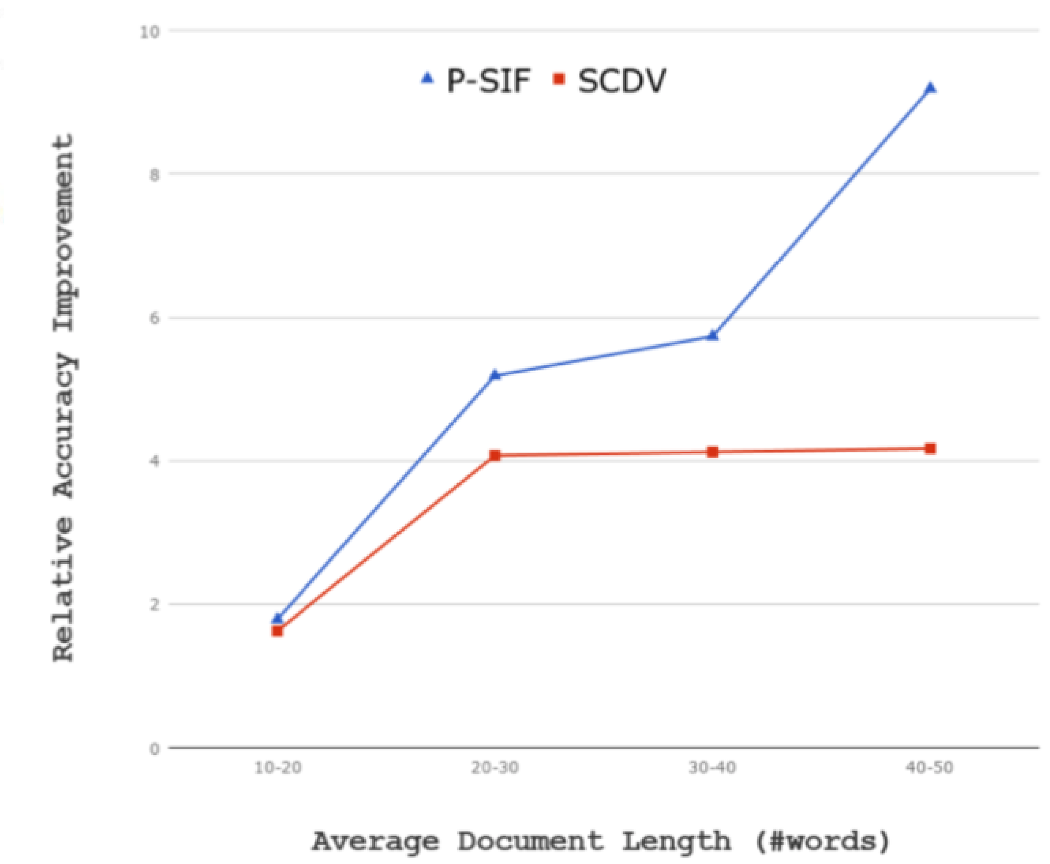
Model	Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)
P-SIF	**86.0**	**86.1**	**86.1**	**86.0**
SCDV	84.6	84.6	84.5	84.6
BoWV	81.6	81.1	81.1	80.9
weight -Avg (SIF)	81.9	81.7	81.9	81.7
BERT (pr)	84.9	84.9	85.0	85.0
NTSG-1	82.6	82.5	81.9	81.2
TWE-1	81.5	81.2	80.6	80.6
Doc2Vec	75.4	74.9	74.3	74.3

- Multi-label text classification on Reuters

Model	Prec@1 (↑)	Prec@5 (↑)	Coverage (↑)	F1-Score (↑)
P-SIF	**94.92**	**37.98**	**93.97**	**82.87**
SCDV	94.20	36.98	93.52	81.75
BoWV	92.90	36.14	91.84	79.16
weight-Avg (SIF)	89.33	35.04	91.68	71.97
BERT (pr)	93.80	37.00	93.70	81.90
TWE-1	90.91	35.49	91.84	79.16
Doc2Vec	88.78	34.51	88.72	73.68

- Experiment on other datasets are reported in the paper

Long vs Short Documents



Effect of Sparse Partitioning

1. Better handling of the multi-sense words
2. Obtains more diverse non-redundant partitions
3. Effectively combine local and global semantics

Takeaways

1. Partition Averaging is better than Averaging
2. Disambiguating multi-sense ambiguity helps
3. Noise in word representations is of huge impact

Limitations

1. Doesn't account for syntax, grammar, and order
2. Disjoint process of partitioning, averaging and task learning

References

- Arora, Sanjeev, et al. *Linear algebraic structure of word senses, with applications to polysemy.* TACL 2018.
- Arora, Sanjeev, et al. *A latent variable model approach to pmi-based word embeddings.* TACL 2016.
- Arora, Sanjeev, et al. *A simple but tough-to-beat baseline for sentence embeddings.* ICLR 2017.