P-SIF: Document Embeddings Using Partition Averaging

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

[1]University of Utah, USA; [2]Microsoft Research, India
[3]InfoEdge Ltd., India
[4]Indian Institute of Technology, Kanpur
[5]Indian Institute of Science, Bangalore

11 February 2020

AAAI 2020, New York

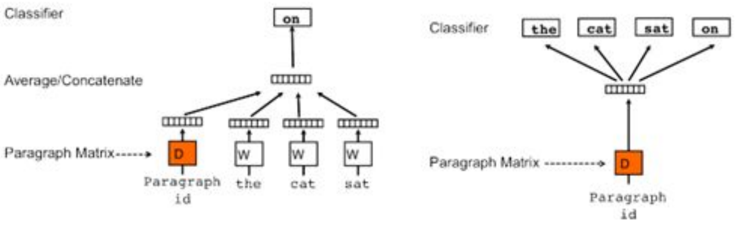
Motivation

- Natural language requires good semantic representations of **textual documents**
  - Text Categorization
  - Information Retrieval
  - Text Similarity

- Good semantic representation of words exists, i.e., **Word2vec (SGNS, CBOW)** created by Mikolov et al., **Glove** (Socher et al.) and many more.

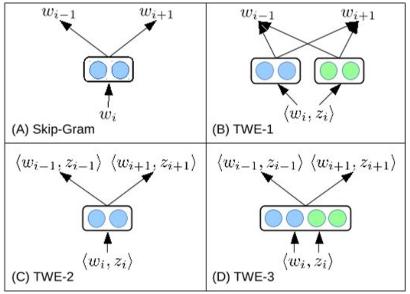
- **What About Documents?**
  - **Multiple Approaches** based on **local context, topic modelling, context sensitive learning**
  - **Semantic Composition** in natural language is the task of modelling the meaning of a larger piece of text *(document)* by composing the meaning of its constituents/parts *(words).*
    - *Our work focus on using simple semantic composition*

Efforts for Document Representation

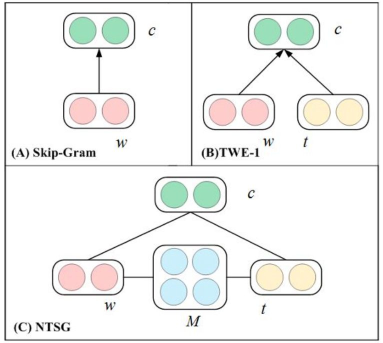


Doc2Vec (Le & Mikolov, 2014)
Local + Global context

Deep Learning
LSTM, RNN, Bi-LSTM,
RTNN, LSTM Attention
Contextual Embedding
ELMo, BERT



TWE (Liu et al., 2015a)
Topic Modelling



~~Larger~~
~~Document~~
~~Multiple~~
~~topic~~

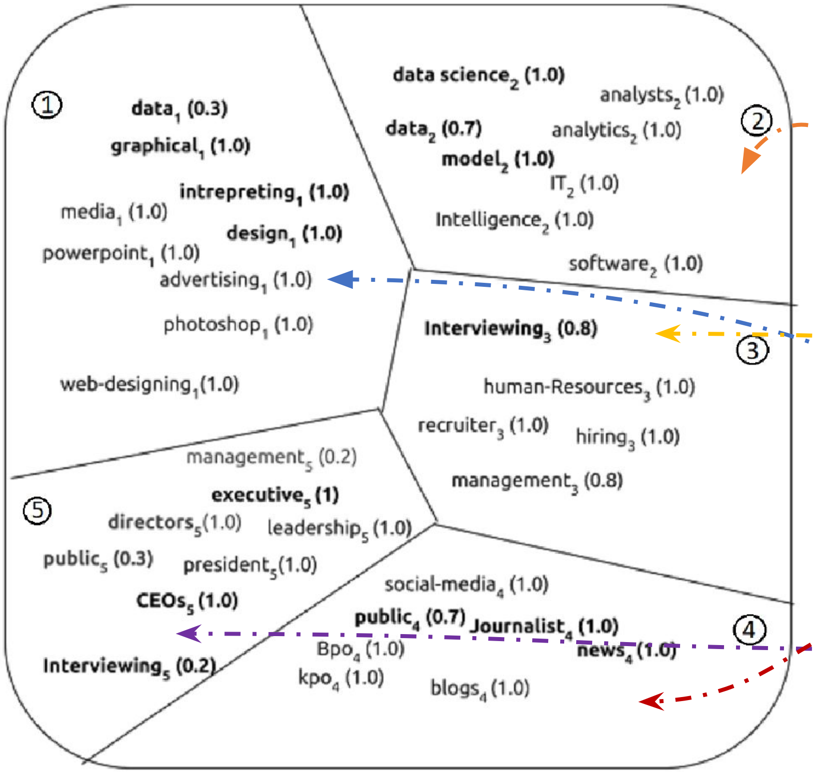
Sentence
Embedding

Graded Weighted M
2015, Arora
Weighted Average
position

NTSG (Liu et al., 2015b)
Topic Modelling + Context Sensitive Learning

Averaging vs Partition Averaging

"Data journalists deliver the news of data science to general public, they often take part in interpreting the data models, creating graphical designs and interviewing the director and CEOs."



SIMPLE AVERAGING

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

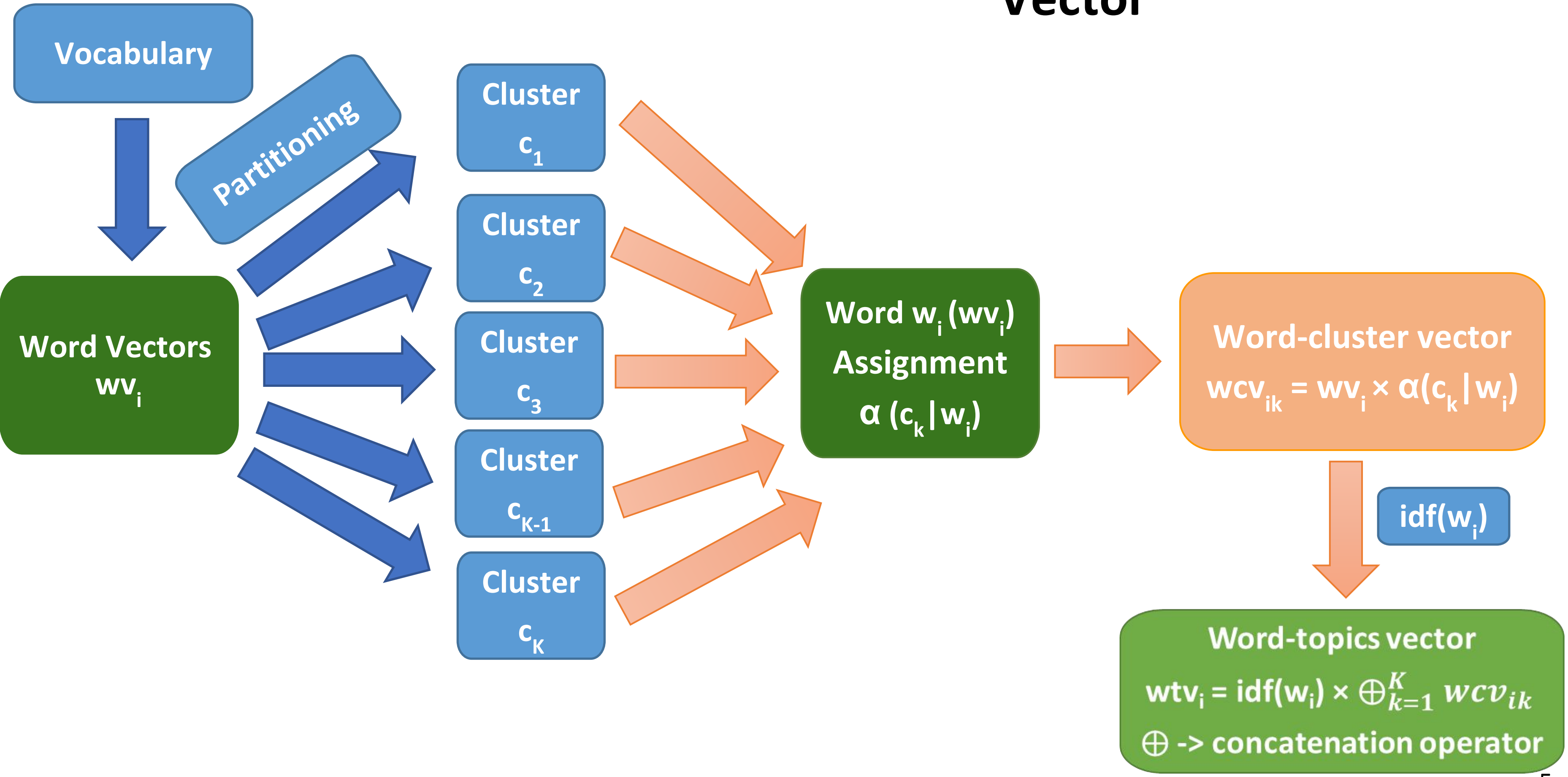
Here, $\oplus$ represent concatenation, and + represent addition

WEIGHTED PARTITION AVERAGING

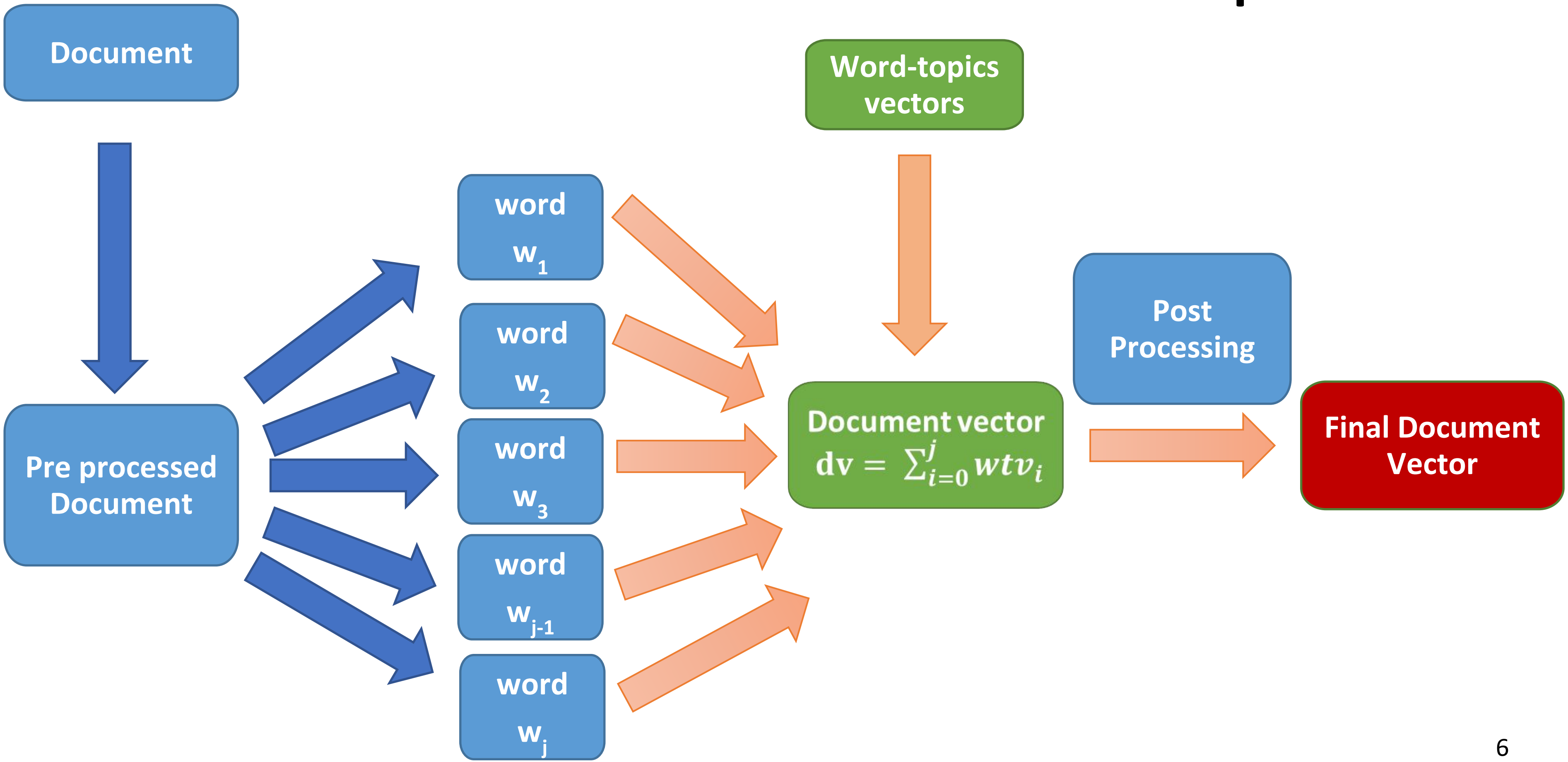
$$(\vec{v}_{\text{interpreting}_1} + \vec{v}_{\text{graphical}_1} + \vec{v}_{\text{design}_1} + 0.3 * \vec{v}_{\text{data}_1}) \oplus (0.7 * \vec{v}_{\text{data}_2} + \vec{v}_{\text{datascience}_2} + \vec{v}_{\text{models}_2}) \oplus (\vec{v}_{\text{journalist}_4} + \vec{v}_{\text{news}_4} + 0.7 * \vec{v}_{\text{public}_4})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

+ within a partition and $\oplus$ across partitions

Pre-computation of Word-topics Vector



Final Document Representation



Connection with simple weighted averaging

Similar to simple weighted averaging model
we average **word topic vectors** instead of **word vectors**

Ways to Partition Vocabulary

Hard Clustering: Assign each word to a single cluster. K-means over word vectors.

Soft Clustering: Assign each word to multiple cluster with probability. Gaussian Mixture Model (GMM) over word vectors

Soft Clustering + Thresholding: Soft Clustering followed by post -processing assignment value below certain threshold (th) to exact 0.

$$\alpha\ (c_k|w) < th \rightarrow \alpha\ (c_k|w) = 0$$

Dictionary Learning: Use sparsity constraint to find minimal basis set. Analogous to soft clustering with sparsity constraint (only k/K non-zero). K-svd over word vectors.

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

SGNS: word2vec algorithm namely Skip Gram with Negative Sampling. Give uni-sense embedding per words.

Doc2VecC: Like SGNS give uni-sense embedding per word but train with corruptions in examples this encourse zeroing of common word vectors.

Multi-Sense + Doc2VecC: Annotated each word in corpus with it sense, for e.g. word bank as (bank#1 , bank#2) based on context in use (river bank, financial institution) and then train Doc2VecC on annotated corpus.

BERT: Fine grain context aware representation, shown to capture word order and syntax in sentence.

Ways to Represent Words

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

For effect of using multi-sense embedding see our recent work at ECAI 20, Spain

Multi-Class Classification – 20NewsGroup (40-80 words)

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

Partition Averaging Algorithm

- P-SIF: Dictionary learning
- SCDV (Mekala et. al, EMNLP 17): GMM clustering
- BoWV (Gupta et. al, Coling 16): k-means clustering
- weight-Avg (SIF, Arora et. al. 17): No partitioning

P-SIF uses only 20 partitions for best performance compared to 60 in SCDV

Multi-Class Classification – 20NewsGroup (40-80 words)

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

P-SIF uses only 20 partitions for best performance compared to 60 in SCDV

Multi-Label Classification - Reuters (200-400 words)

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

Partition Averaging Algorithm

- P-SIF: Dictionary learning
- SCDV (Mekala et. al, EMNLP 17): GMM clustering
- BoWV (Gupta et. al, Coling 16): k-means clustering
- weight-Avg (SIF, Arora et. al. 17): No partitioning

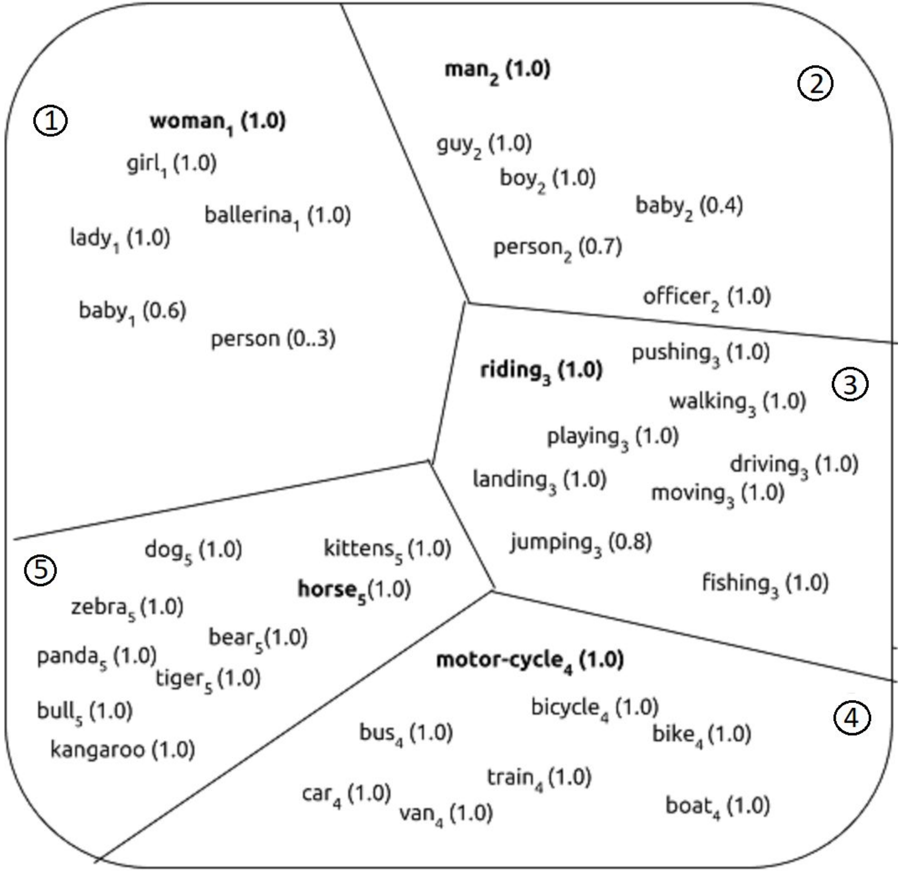
Effect of partitioning more significant than 20NewsGroup due to larger document length

Multi-Label Classification - Reuters (200-400 words)

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

Effect of partitioning more significant than 20NewsGroup due to large length

Semantic Textual Similarity (27 Datasets)



	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**

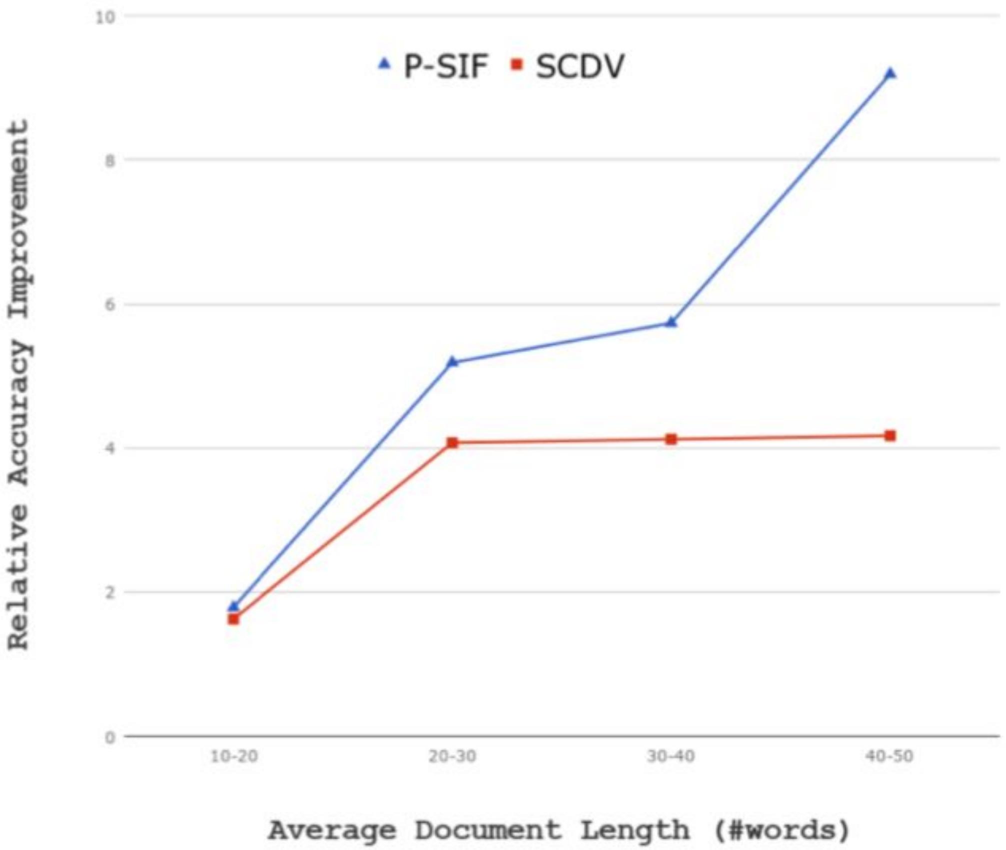
Semantic Textual Similarity (27 Datasets)

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

Results (Pearson r X 100) on Semantic Textual Similarity

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

Relative Performance (P-SIF – SIF)/SIF (%) Improvement



Theoretical Justification

We provide theoretical justifications of P-SIF by showing connections with **random walk-based latent variable models** in (Arora et al. 2016a; 2016b, TACL 16,18) and SIF embedding (Arora, Liang, and Ma 2017, ICLR 17).

We **relax one assumption** and **introduce context jump** in the SIF embedding to show that our approach P-SIF embedding is a **generalization** of the SIF sentence embedding which is a special case of with number of clusters K = 1.

Takeaways

✓ Replace weighted **word vector averaging (SIF)** with **partition based averaging (P-SIF)** for a **strong baseline** for **document representation**. (capture **local + global semantics**)

- **Dictionary Learning** better than **GMM Clustering + Hard Threshold**: Imposing sparsity constraint during partitioning is beneficial .
- **GMM/Dictionary Learning** better than **K-means Clustering** : Soft clustering is better than hard clustering

✓ **Noise in words level representation is influential** on the final downstream tasks. **Doc2VecC** for better word representation than **SGNS**.

Paper ID: 3656, visit our poster in the evening session to know more !
(such as interesting connections to kernels)

my email : keviv9@gmail.com , web: vgupta123.github.io

Acknowledgement

- Anonymous reviewers of ICLR’19 and AAAI’20 whose reviews really helped in improving the paper
- AAAI’20 Student Scholar and Volunteer Program for the needful support
- Prof. Vivek Srikumar, Prof. Ellen Riloff, Prof. Aditya Bhaskara and Prof. Suresh Venkatasubramanian of School of Computing, University of Utah for useful feedback
- Microsoft Research Lab, Bangalore; School of Computing, University of Utah and Indian Institute of Technology, Kanpur for needed support and guidance

References

- **BoWV** : Vivek Gupta and Harish Karnick et al, "*Product Classification in e-Commerce using Distributional Semantics*", In Proc COLING 2016
- **SCDV** : Dheeraj Mekala*,Vivek Gupta*, Bhargavi Paranjape and Harish Karnick, “*Sparse Composite Document Vectors using Soft Clustering over Distributional Semantics*”, In Proc EMNLP 2017
- **SCDV-MS** : Vivek Gupta et. al. “Word Polysemy Aware Document Vector Estimation”, In Proc ECAI 2020.
- **NTSG** : Pengfei Liu and Xipeng Qiu et al., *"Learning Context-Sensitive Word Embedding's with Neural Tensor Skip-Gram Model”*, In Proc IJCAI 2015
- **TWE** : Yang Liu and Zhiyuan Liu et al, *"Topical Word Embeddings”* In Proc AAAI, 2015
- **Lda2Vec** : Chris Moody *"Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec”*, arXiv:1605.02019
- **WMD** : Matt J. Kusner et al., *"From Word Embeddings To Document Distance*”, In ICML 2015
- **WME** : Lingfei Wu, Ian E.H. Yen et. al., *“Word Mover's Embedding: From Word2Vec to Document Embedding”,* In EMNLP 2018
- **SIF** : Sanjeev Arora and Yingyu Liang “*A Simple but tough-to-beat baseline for sentence embedding's*”, In ICLR 2017
- **Polysemy** : Sanjeev Arora and Yuanzhi Li et al. “*Linear algebraic structure of word senses, with applications to polysemy”,* In TACL 2018
- **Doc2vec** *:* Quoc V Le and Tomas Mikolov. *“Distributed Representations of Sentences andDocuments"* In: ICML 2014

Limitations

✘ Doesn't account for syntax, grammar, and words order and only focuses on effective capturing of local and global semantics.

✘ Currently, a disjoint process of partitioning, averaging and task learning: can we model everything as a single joint process?

Positive Qualitative Results (MSRvid)

sentence1	sentence2	GT	NGT	SIF_{sc}	P-SIF_{sc}
People are playing baseball .	The cricket player hit the ball .	0.5	0.1	0.2928	0.0973
A woman is carrying a boy .	A woman is carrying her baby .	2.333	0.4666	0.5743	0.4683
A man is riding a motorcycle .	A woman is riding a horse .	0.75	0.15	0.5655	0.157
A woman slices a lemon .	A man is talking into a microphone .	0	0	-0.1101	-0.0027
A man is hugging someone .	A man is taking a picture .	0.4	0.08	0.2021	0.0767
A woman is dancing .	A woman plays the clarinet .	0.8	0.16	0.3539	0.1653
A train is moving .	A man is doing yoga .	0	0	0.1674	-0.0051
Runners race around a track .	Runners compete in a race .	3.2	0.64	0.7653	0.6438
A man is driving a car .	A man is riding a horse .	1.2	0.24	0.3584	0.2443
A man is playing a guitar .	A woman is riding a horse .	0.5	0.1	-0.0208	0.0955
A man is riding on a horse .	A girl is riding a horse .	2.6	0.52	0.6933	0.5082
A woman is deboning a fish .	A man catches a fish .	1.25	0.25	0.4538	0.2336
A man is playing a guitar .	A man is eating pasta .	0.533	0.1066	-0.0158	0.0962
A woman is dancing .	A man is eating .	0.143	0.0286	-0.1001	0.0412
The ballerina is dancing .	A man is dancing .	1.75	0.35	0.512	0.3317
A woman plays the guitar .	A man sings and plays the guitar .	1.75	0.35	0.5036	0.3683
A girl is styling her hair .	A girl is brushing her hair .	2.5	0.5	0.7192	0.5303
A guy is playing hackysack	A man is playing a key-board .	1	0.2	0.3718	0.2268
A man is riding a bicycle .	A monkey is riding a bike .	2	0.4	0.6891	0.4614
A woman is swimming underwater .	A man is slicing some carrots .	0	0	-0.2158	-0.0562
A plane is landing .	A animated airplane is landing .	2.8	0.56	0.801	0.6338
The missile exploded .	A rocket exploded .	3.2	0.64	0.8157	0.6961
A woman is peeling a potato .	A woman is peeling an apple .	2	0.4	0.6938	0.5482
A woman is writing .	A woman is swimming .	0.5	0.1	0.3595	0.2334
A man is riding a bike .	A man is riding on a horse .	2	0.4	0.6781	0.564
A panda is climbing .	A man is climbing a rope .	1.6	0.32	0.4274	0.3131
A man is shooting a gun .	A man is spitting .	0	0	0.2348	0.1305

Negative Qualitative Results (MSRvid)

sentence1	sentence2	GT	NGT	SIF_{sc}	$P\text{-}SIF_{sc}$
takes off his sunglasses .	A boy is screaming .	0.5	0.1	0.1971	0.3944
The rhino grazed on the grass .	A rhino is grazing in a field .	4	0.8	0.7275	0.538
An animal is biting a persons finger .	A slow loris is biting a persons finger .	3	0.6	0.6018	0.7702
Animals are playing in water .	Two men are playing ping pong .	0	0	0.0706	0.2238
Someone is feeding a animal .	Someone is playing a piano .	0	0	-0.0037	0.1546
The lady sliced a tomatoe .	Someone is cutting a tomato .	4	0.8	0.693	0.5591
The lady peeled the potatoe .	A woman is peeling a potato .	4.75	0.95	0.7167	0.5925
A man is slicing something .	A man is slicing a bun .	3	0.6	0.5976	0.4814
A boy is crawling into a dog house .	A boy is playing a wooden flute .	0.75	0.15	0.1481	0.2674
A man and woman are talking .	A man and woman is eating .	1.6	0.32	0.3574	0.4711
A man is cutting a potato .	A woman plays an electric guitar .	0.083	0.0166	-0.1007	-0.2128
A person is cutting a meat .	A person riding a mechanical bull	0	0	0.0152	0.1242
A woman is playing the flute .	A man is playing the guitar .	1	0.2	0.1942	0.0876

Kernel Connection with Embeddings

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

word vector averaging

Topical Word Embedding (TWE)

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

Our Partitioning Model (P-SIF)

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

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

word mover distance