Document Vector Estimation using Partition Word-Vectors Averaging

Vivek Gupta
PhD Student
School of Computing
University of Utah

Microsoft Research

20 June 2019

IBM Intern Highlight Talk

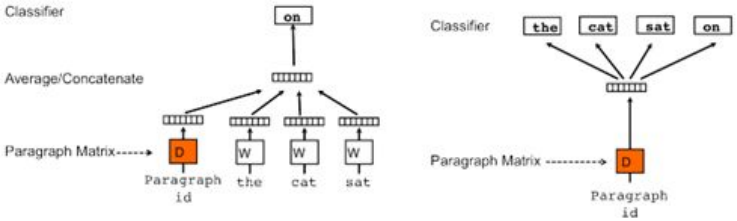
Motivation

- 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.

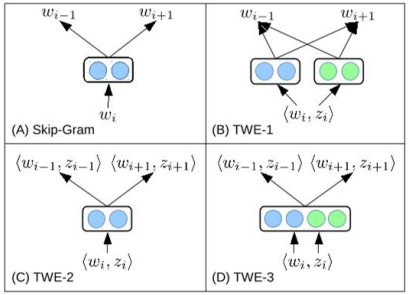
- **What About Documents?**
  - **Multiple Approaches** based on **local context, topic modelling, context sensitive learning**
  - **Semantic Composition** in natural language is the task of modelling the meaning of a larger piece of text *(document)* by composing the meaning of its constituents/parts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

Efforts for Document Representation



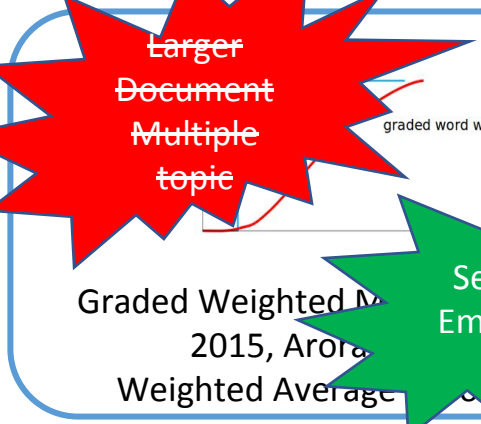
Doc2Vec (Le & Mikolov, 2014)
Local + Global context

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



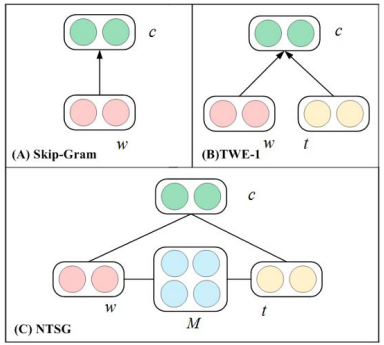
TWE (Liu et al., 2015a)
Topic Modelling

~~Larger Document Multiple topic~~



Graded Weighted M
2015, Arora
Weighted Average position

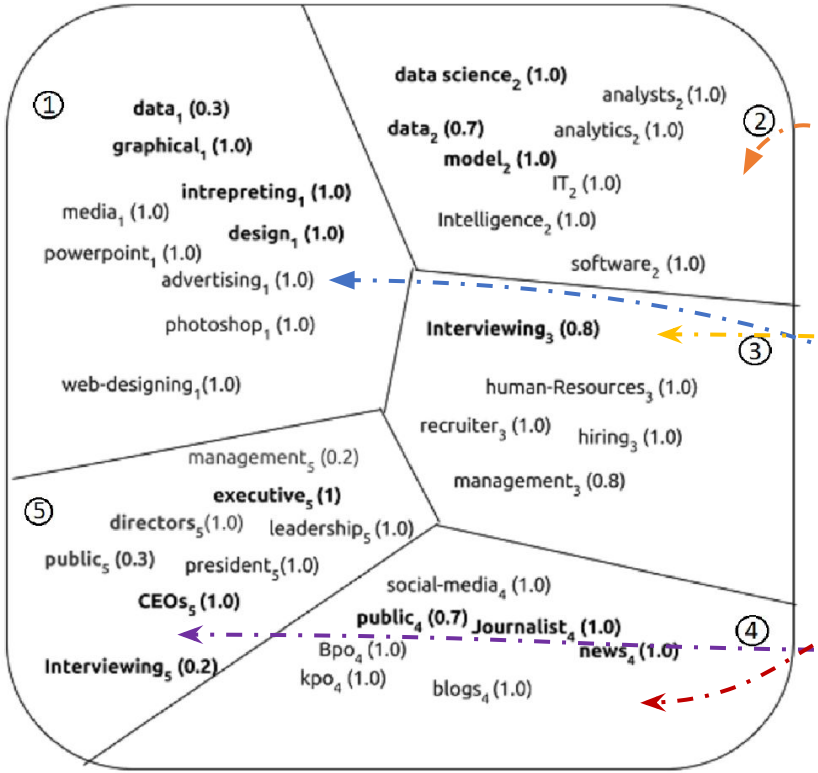
Sentence Embedding



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 CEO's."



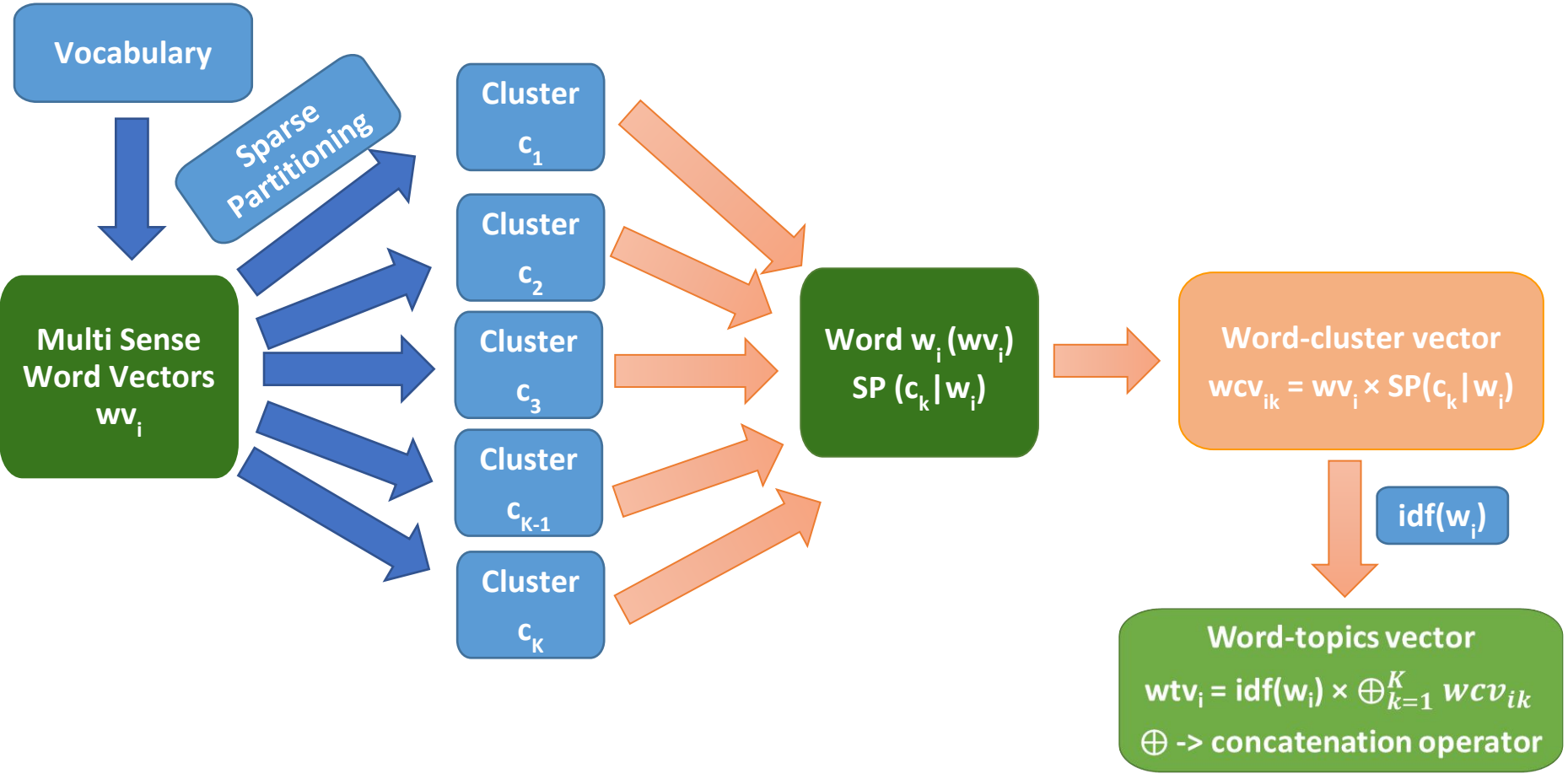
SIMPLE AVERAGING

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

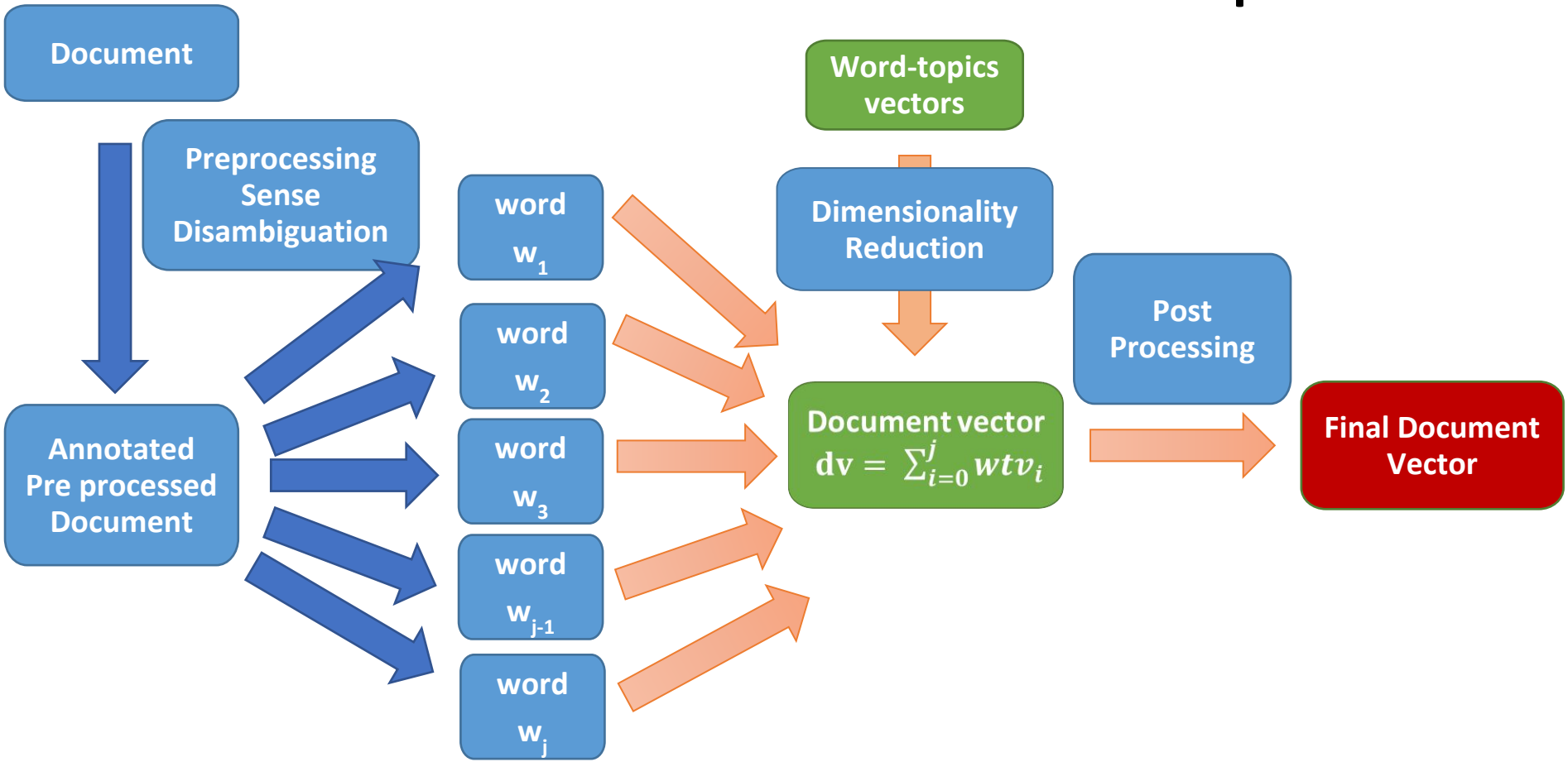
WEIGHTED PARTITION AVERAGING

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

Pre-computation of word-topics vector



Final Document Representation



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

Nice Connection

Ways to Partition Vocabulary

Name	Partition Type	Properties	Method
K-Means	Hard Clustering	Polysemic Words ☹,Vectors Sparsity 😊, Partition Diversity ☹ ,Document Sparsity😊 Pre-Computation ☹	BoWV [1] Coling'16
GMM	Fuzzy Clustering	Polysemic Words😊, Vectors Sparsity☹, Partition Diversity☹, Document Sparsity😊, Pre-Computation 😊	SCDV [2] EMNLP'17
Sparse GMM	Sparse Fuzzy Clustering	Polysemic Words 😊, Vectors Sparsity😊, Partition Diversity☹, Document Sparsity😊,Pre-Computation 😊	SCDV-MS [3] NAACL'19
K-SVD	Dictionary Learning	Polysemic Words 😊, Vectors Sparsity😊, Partition Diversity 😊, Document Sparsity😊,Pre-Computation 😊	P-SIF [4] under review

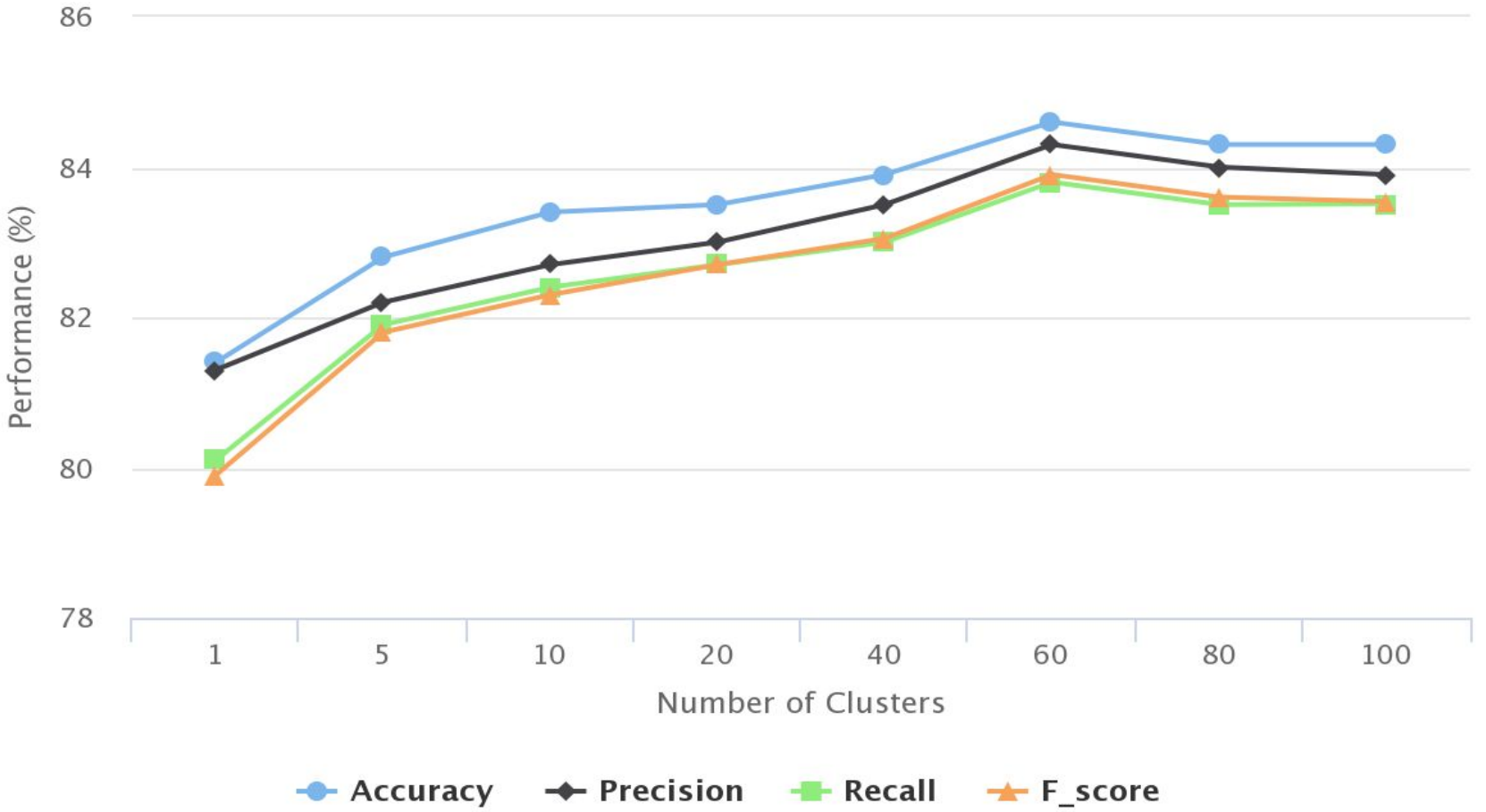
Ways to represent words

Name	Embedding Type	Properties	Method
SGNS	Noise Sensitive	Common word removal☹, Context Sensitive (multi-sense)☹ Capture Order-Syntax Information ☹	BoWV SCDV
Doc2VecC	Noise Insensitive	Common word removal😊, Context Sensitive (multi-sense)☹ Capture Order-Syntax Information ☹	SCDV-MS P-SIF
AdaGram + Doc2VecC	Multi-Sense Noise Insensitive	Common word removal😊, Context Sensitive (multi-sense)😊, Capture Order-Syntax Information ☹	SCDV-MS
Elmo/BERT	Contextual Syntax Preserving	Common word removal😊,Context Sensitive (multi-sense)😊, Capture Order-Syntax Information 😊	Yet to Explore

Ways to weight words

Technique	Operation	Method
Inverse document frequency	Concatenation	BOWV
Inverse document frequency	Multiplication	SCDV
Smooth Inverse frequency	Multiplication	P-SIF

Effect of Partitioning (Text Categorization 20NewsGroup)



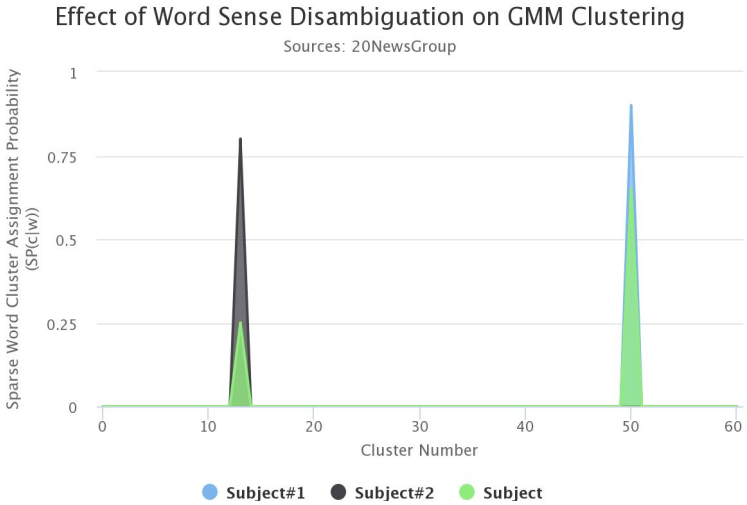
Topic Modelling (Coherence) using GMM

GMM	LTSG	LDA
-85.23	-92.33	-108.72

Topic Image			Topic Health			Topic Mail		
GMM	LTSG	LDA	GMM	LTSG	LDA	GMM	LTSG	LDA
file	image	image	heath	stimulation	doctor	ftp	anonymous	list
bit	jpeg	file	study	diseases	disease	mail	faq	mail
image	gif	color	medical	disease	coupons	internet	send	information
files	format	gif	drug	toxin	treatment	phone	ftp	internet
color	file	jpeg	test	toxic	pain	email	mailing	send
format	files	file	drugs	newsletter	medical	send	server	posting
images	convert	format	studies	staff	day	opinions	mail	email
jpeg	color	bit	disease	volume	microorganism	fax	alt	group
gif	formats	images	education	heaths	medicine	address	archive	news
program	images	quality	age	aids	body	box	email	anonymous
-67.16	-75.66	-88.79	-66.91	-96.98	-100.39	-77.47	-78.23	-95.47

Context Sensitive Learning (Multi-Sense Effect)

Sentence (Context Words)	Prominent Cluster Words
The math **subject** is a nightmare for many students	physics, chemistry, math, science
In anxiety, he sent an email without a **subject**	mail, letter, email, gmail
After promotion, he went to Maldives for spring **break**	vacation, holiday, trip, spring
Breaking government websites is common for hackers	encryption, cipher, security, privacy
Use **break** to stop growing recursion loops	if, elseif, endif, loop, continue
The S.I. **unit** of distance is meter	calculation, distance, mass, length
Multimeter shows a **unit** of 5V	electronics, KWH, digital, signal
His **interest** lies in astrophysics	information, enthusiasm, question
Banks **interest** rates are controlled by RBI	bank, market, finance, investment



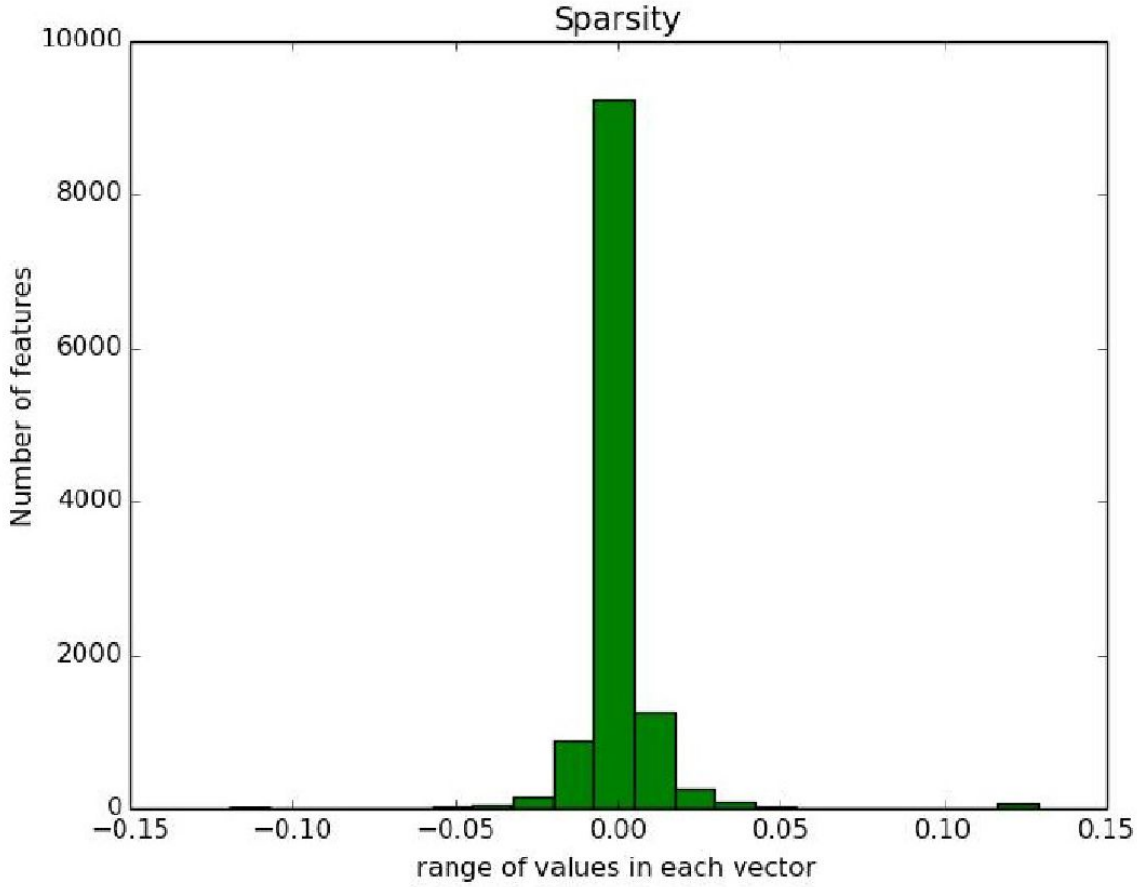
Black is the *chance of "subject#1" (multisense)* belonging to cluster #14

Blue is the *chance of "subject#2" (multisense)* belonging to cluster #50

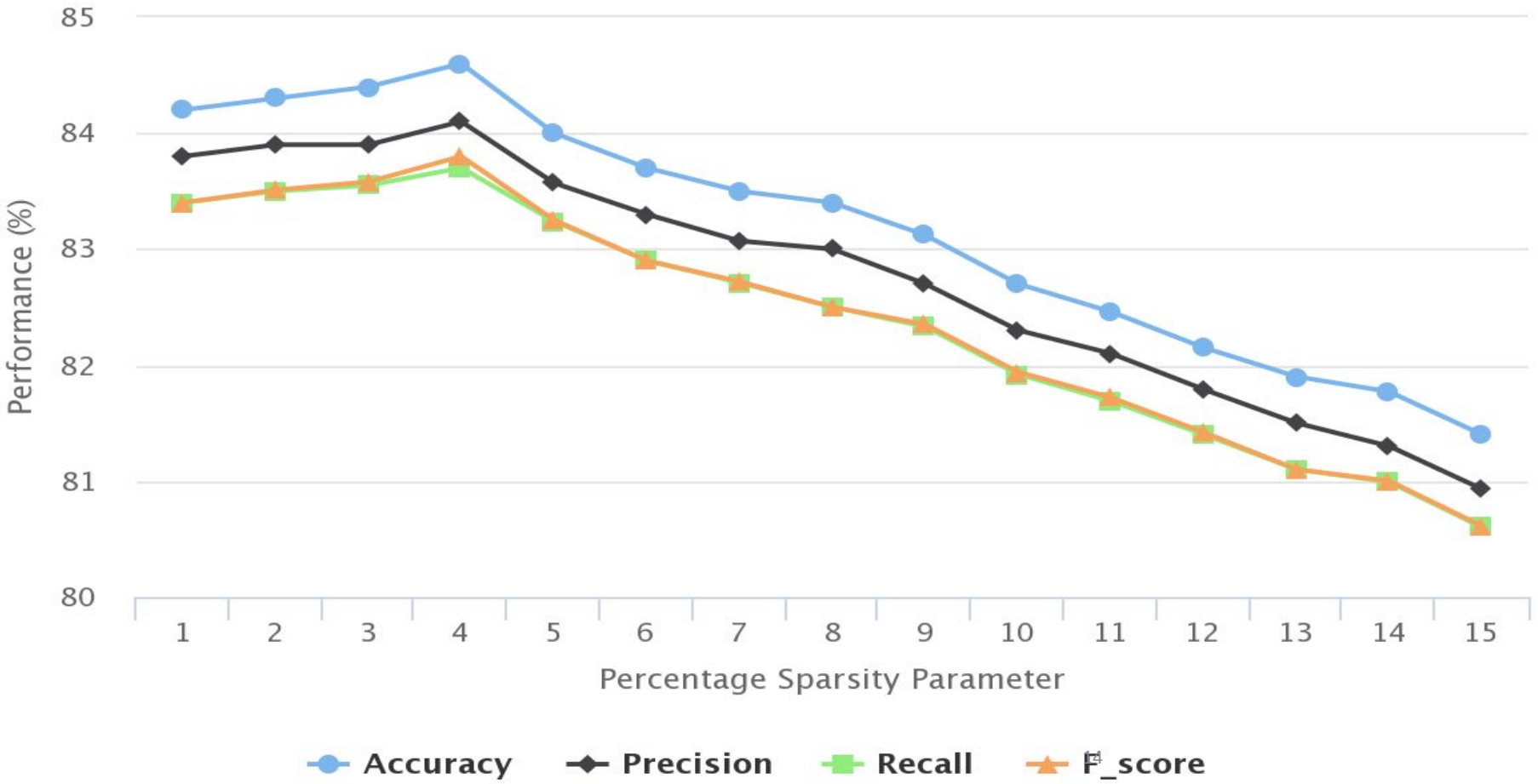
Green is the *chance of "subject" (unisense)* belonging to cluster #14 & #50

All other *chances* for other clusters are negligibly small

Sparsity in Representation



Effect of Sparsity (SCDV) (Text Categorization 20NG)



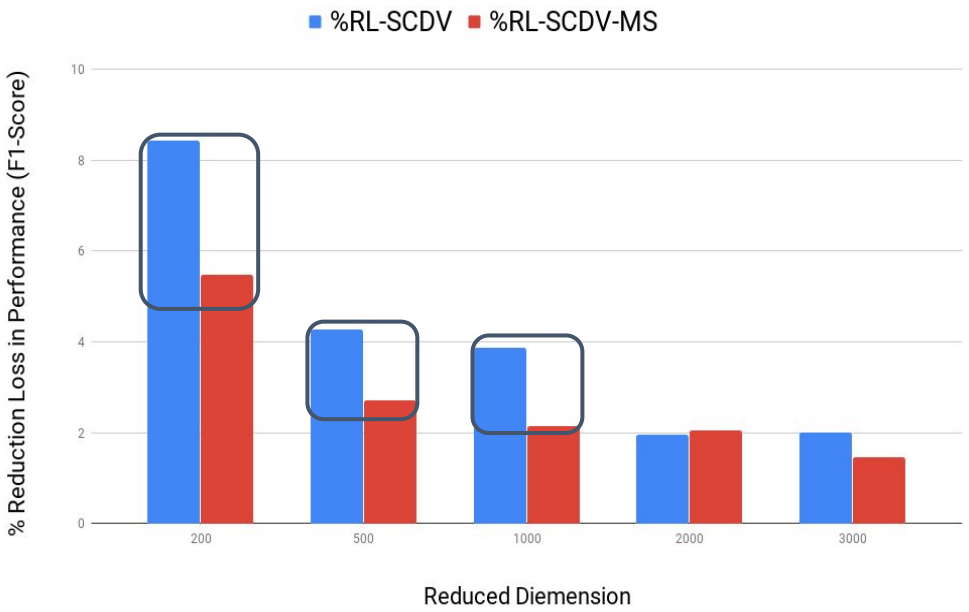
Lower Dimension Embedding (Sparsity Effect)

Manifold Learning Algorithms

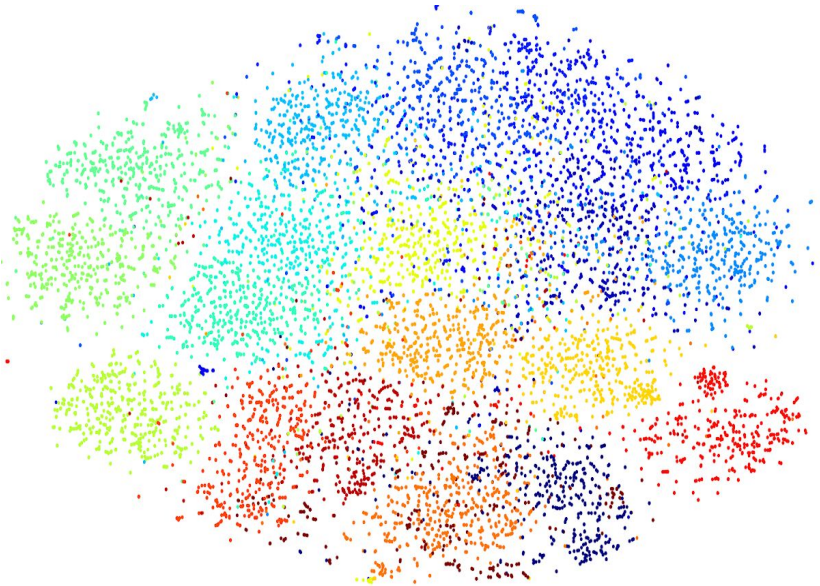
- PCA-Subspace
- Random Projection
- **Autoencoders**

All work effectively well

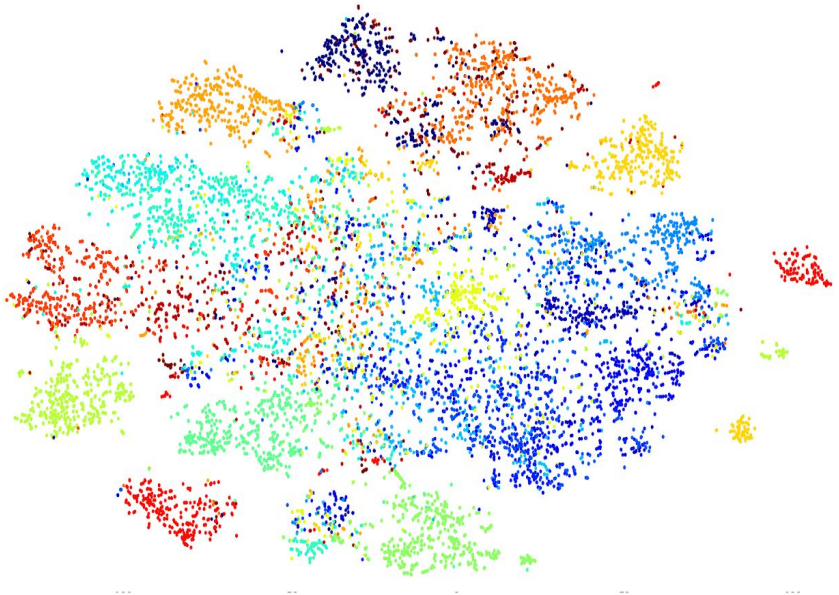
Better reduction compared to the SCDV-word-topic-vector for all methods



t-sne visualization (better separation)



Doc2vec



SCDV

Multi-Class Classification - 20NewsGroup Dataset

Multi-Class Classification- 20 NewsGroup – 20 classes, Equal Sampling, 200-300 word documents, Language: English

Model	Accuracy	Precision	Recall	F1-Score
SCDV-MS	86.2	86.2	86.2	86.2
R-SCDV-MS	84.9	84.9	84.9	84.9
SCDV	84.6	84.6	84.5	84.6
BoE	83.1	83.1	83.1	83.1
BoWV	81.6	81.1	81.1	80.9
NTSG-1	82.6	82.5	81.9	81.2
LTSG	82.8	82.4	81.8	81.8
TWE-1	81.5	81.2	80.6	80.6
PV-DBoW	75.4	74.9	74.3	74.3
PV-DM	72.4	72.1	71.5	71.5

17

Multi-Label Classification - Reuters Dataset

Multi-Label Classification- Reuters - ~5000 labels, Unequal Sampling, 400-500 word documents, Language: English

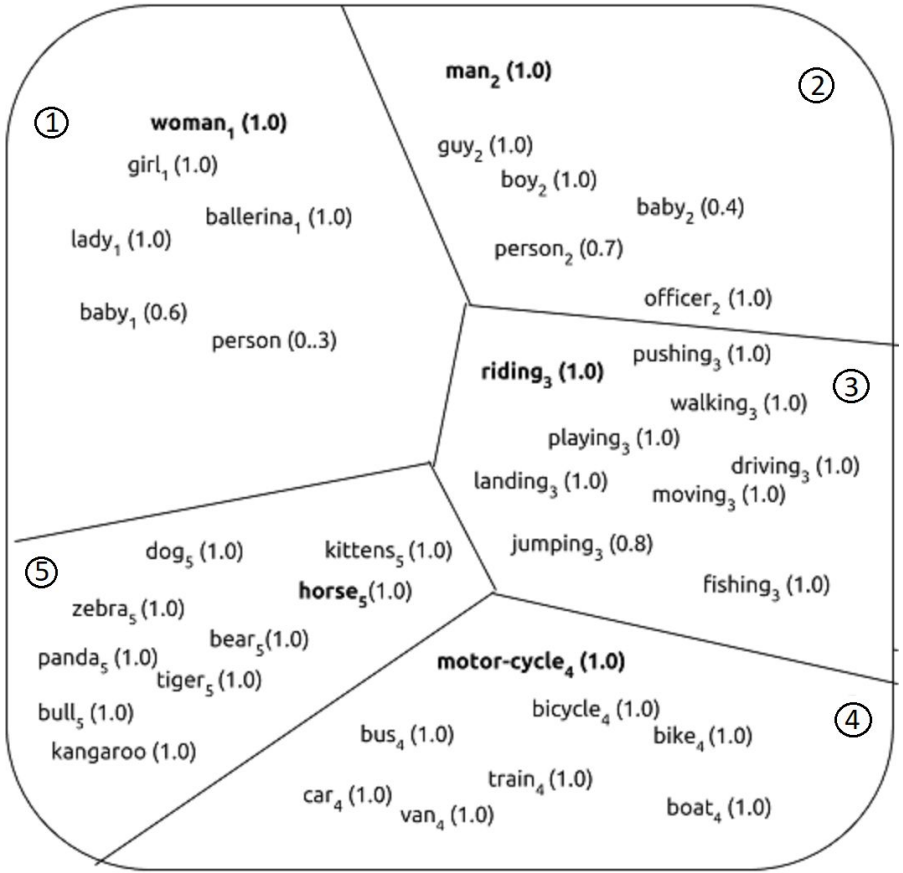
Model	Prec@1 nDCG@1	Prec@5	nDCG@5	Coverage	LRAPS	F1-Score
SCDV-MS	95.06	37.56	50.20	94.13	94.21	82.71
R-SCDV-MS	93.56	37.00	49.47	93.61	92.96	81.94
SCDV	94.20	36.98	49.55	93.52	93.30	81.75
BoWV	92.90	36.14	48.55	91.84	91.46	79.16
TWE-1	90.91	35.49	47.54	91.84	90.97	79.16
PV-DBoW	88.78	34.51	46.42	88.72	87.43	73.68
PV-DM	87.54	33.24	44.21	86.85	86.21	70.24

Ablation and Efficiency Analysis

Ablation	20NewsGroup	Reuters
w/o Sparsity	85.78 ± 0.002	82.17 ± 0.001
w/o Doc2VecC	85.41 ± 0.001	82.08 ± 0.002
w/o Multi-Sense	85.16 ± 0.001	82.43 ± 0.001
w/o All	84.61 ± 0.004	81.77 ± 0.003
w All	86.16 ± 0.002	82.71 ± 0.002

Method	Vocab	$\vec{wtv}$ Dim	$\vec{wtv}$ Sparsity (%)	Cluster Time (sec)	Feature Time (μ sec)	Predict (μ sec)	Training Time (min)	Model Size (KB)	$\vec{wtv}$ Space (MB)
SCDV	15591	12000	1	**242**	2.56	119	82	1900	748
SCDV-MS	25466	12000	**98**	569	**0.06**	111	79	1900	**71**
R-SCDV-MS	25466	2000	0	576	0.86	**14**	**66**	**333**	203

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	$\vec{v}_{\text{zero}_1} \oplus \vec{v}_{\text{man}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{motorcycle}_4} \oplus \vec{v}_{\text{zero}_5}$

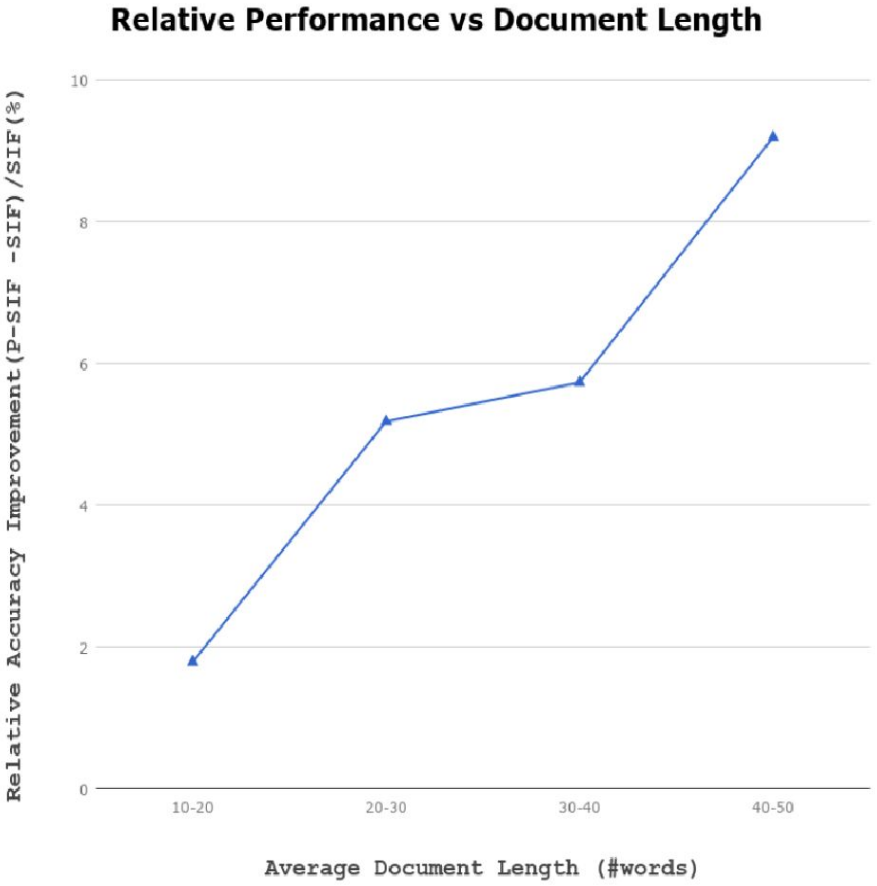
	Document 2 (d_n^2)
Doc	A woman is riding a horse
SIF	$\vec{v}_{\text{woman}_1} + \vec{v}_{\text{riding}_3} + \vec{v}_{\text{horse}_5}$
P-SIF	$\vec{v}_{\text{women}_1} \oplus \vec{v}_{\text{zero}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{zero}_4} \oplus \vec{v}_{\text{horse}_5}$

SIMILARITY SCORES		
Ground Truth	SIF	P-SIF
0.15	0.57	0.16

Results (Pearson r X 100) on Semantic Textual Similarity

Supervised or Not	Supervised								UnSupervised			Semi Supervised			Our
Tasks	**PP**	**PP -Proj**	**DAN**	**RNN**	**iRNN**	**LSTM (no)**	**LSTM (o.g.)**	**GRAN**	**ST**	**avg Glove**	**tfidf Glove**	**avg -PSL**	**Glove +WR**	**PSL +WR**	**Our +PSL**
STS12	58.7	60.0	56.0	48.1	58.4	51.0	46.4	62.5	30.8	52.5	58.7	52.8	56.2	59.5	65.7
STS13	55.8	56.8	54.2	44.7	56.7	45.2	41.5	63.4	24.8	42.3	52.1	46.4	56.6	61.8	64.0
STS14	70.9	71.3	69.5	57.7	70.9	59.8	51.5	75.9	31.4	54.2	63.8	59.5	68.5	73.5	74.8
STS15	75.8	74.8	72.7	57.2	75.6	63.9	56.0	77.7	31.0	52.7	60.6	60.0	71.7	76.3	77.3
SICK14	71.6	71.6	70.7	61.2	71.2	63.9	59.0	72.9	49.8	65.9	69.4	66.4	72.2	72.9	73.4
Twitter15	52.9	52.8	53.7	45.1	52.9	47.6	36.1	50.2	24.7	30.3	33.8	36.3	48.0	49.0	54.9

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



Theoretical Justification

- We showed connections of P-SIF with generative random-walk based latent variable models (Arora et. al. 2016a)
- Total number of topics in entire corpus (K) and can be determine by sparse dictionary learning (Arora et. al. 2016b)
- The context vector does not change significantly much while words are generated from random walk except topic change
- The partition function remain same in all directions for only words coming from a same context
- Taylor expansion followed by Maximum Likelihood Estimation over the distribution give the required context vector.
- Concatenation of context vector give the required document embedding.

Kernel Connection of 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

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

- Our Partitioning Model

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

$$K^3(D_A, D_B) = \frac{1}{n}\sum_{i=1}^{n}\max_j\langle \vec{wv}_{w_i^A} \cdot \vec{wv}_{w_j^B}\rangle$$

- Relax word mover distance

$$K^5(D_A, D_B) = K^3(D_A, D_B) + K^4(D_A, D_B)$$

- Word mover distance

Takeaways

✓ Instead of using a simple weighted word vector averaging a partition based weighted average could be a stronger baseline for document representation.

✓ Composition operation like addition and concatenation can have huge impact on the document representation.

✓ Disambiguating multi-sense of words based on context in used (surrounding words) can lead to better document representation.

✓ Sparsity in representation could be useful for learning a lower level representation manifold efficiently.

✓ Noise in words level representation can have huge impact on the final downstream tasks.

Limitations

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

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

Product Impact

✓Flipkart e-Commerce seller platform used a faster distributed version of our approach.

✓One of ensemble feature (with modification) of bing duplicate add detection algorithm.

References

[1] **Sparse Composite Document Vectors using soft clustering over distributional representations,** Dheeraj Mekala*, **Vivek Gupta***, Bhargavi Paranjape, Harish Karnick; Published at EMNLP 2017. [Paper] [PPT]

[2] **Word Polysemy Aware Document Vector Estimation; Vivek Gupta***, Ankit Saw*, Harshit Gupta, Pegah Nokhiz and Partha Talukdar; Presented at NAACL-SRW 2019 (non-archival). email me on my Gmail id for paper and code. [PPT]

[3] **Unsupervised Document Representation using Partition word vector averaging; Vivek Gupta***, Ankit Saw*, Partha Talukdar, and Praneeth Netrapalli; under review: email me on my Gmail id for updated paper and code. [ArXiv] [PPT]

Acknowledgement

Students: Ankit Kumar Saw[2,3](IITKgp)*, Dheeraj Mekala[1]*(IITK), Harshit Gupta[2](IITG), Pegah Nokhiz[2](UoU), Bhargavi Paranjape[1](CMU)

Microsoft Researcher: Praneeth Netrapalli[3](MSR mentor)

Academic Professors: Harish Karnick (IITK, MS Advisor[1]), Partha Pratim Talukdar (IISC-Bangalore[3,4]), Piyush Rai (IITK[4])

*Major Contributors

Thanks for Listening

Questions?

email: keviv9@gmail.com *(Pref)*

homepage: vgupta123.github.io

Blogpost: vivgupt.blogpost.com

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-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