Unsupervised Contextualized Document Representation

Ankur Gupta[1], Vivek Gupta[2]

[1] Indian Institute of Technology Kanpur, [2] School of Computing, University of Utah

1. Document Contextualization

Raw Docs

Messi scored a penalty!
Judge passed the order of ...
The court issued a penalty ....

Contextualized Docs

Messi scored a **penalty$1**!
Judge passed the order of ...
The **court$1** issued a **penalty$0** ....

- Cosine Sim. between embedding of the bold word:

Word	Sentence	Score
Subject	The math **subject1** is difficult He sent the mail without **subject2**	0.71
Apple	The stocks of **apple1** have increased I eat an **apple2** everyday	0.67

- K-Means algorithm to cluster all contextualized representations of all occurrence of the word.
- Word Sense Disambiguation:vocabulary distribution

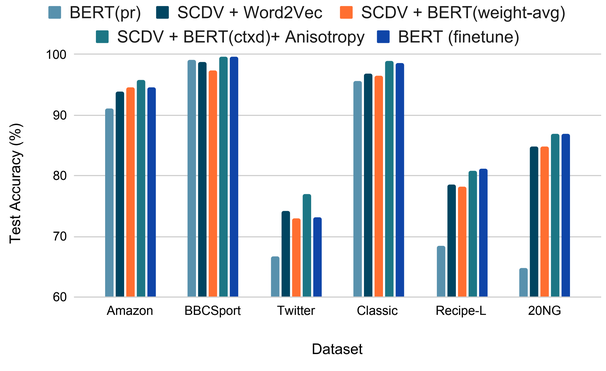
Dataset	k=1	k=2	k≥3
20NG	80.29	13.58	6.23
Amazon	76.12	17.68	6.20
Twitter	80.79	15.60	3.61
BBCSport	87.29	11.56	1.15
Classic	73.63	17.01	9.36
Recipe-l	67.11	13.98	18.91

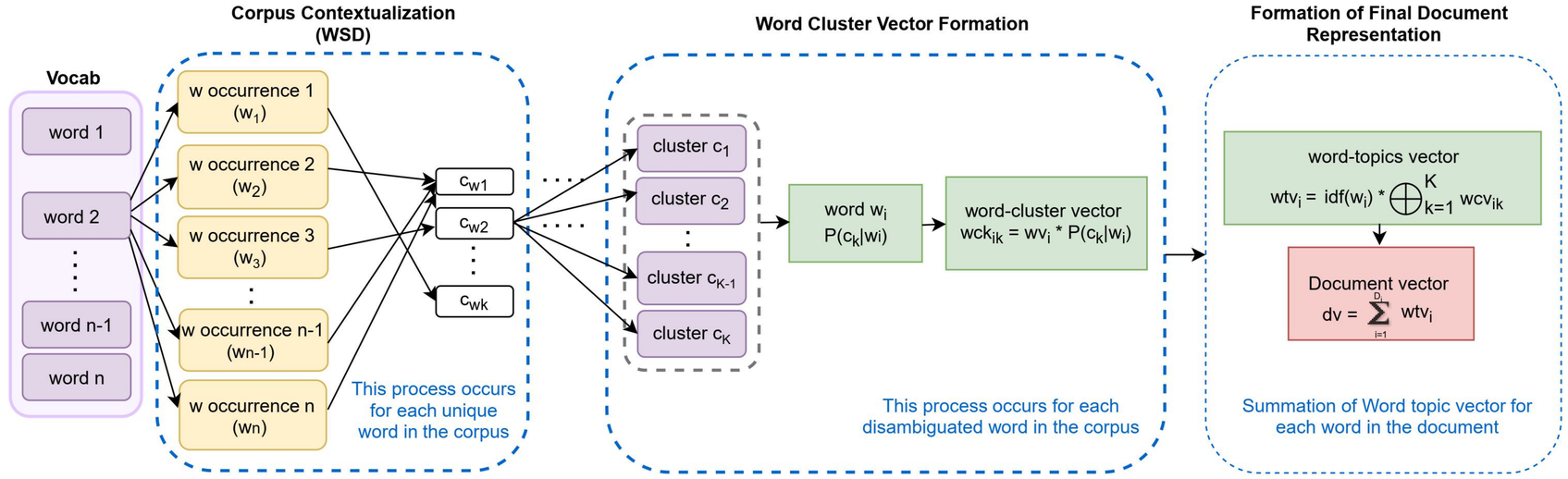
2. Text Classification

- Multi class text classification on 20NewsGroup

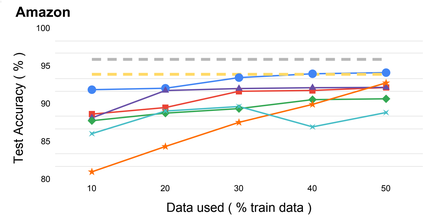
Model	Accuracy	Precision	Recall	F1
SCDV+BERT(cxtd) + Anisotropy	**86.9**	**86.4**	**86.1**	**86.3**
SCDV	84.6	84.6	84.5	84.6
BoWV	81.6	81.1	81.1	80.9
weight -Avg (SIF)	81.9	81.7	81.9	81.7
BERT (pr)	84.9	84.9	85.0	85.0
Doc2Vec	75.4	74.9	74.3	74.3

- Result on various datasets:

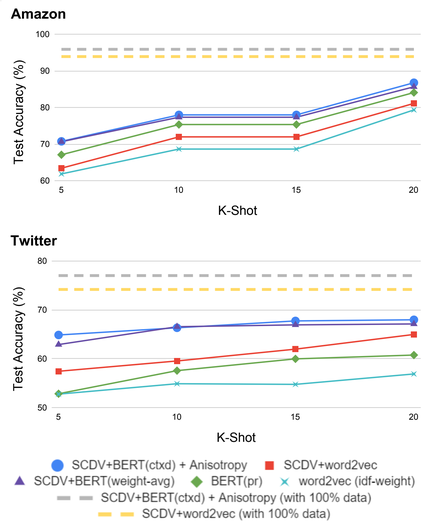




2.1 Low Resource Setting



2.2 Few Shot Setting



3. Text Similarity Task

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

Embedding	Y12	Y13	Y14	Y15	Y16	Avg.
ELMO orig+all	55	51	63	69	64	60.4
ELMO orig+top	54	49	62	67	63	59
BERT(pr)	53	67	62	73	67	64.4
USE	65	**68**	64	77	73	69.4
p-mean	54	52	63	66	67	60.4
fastText	58	58	65	68	64	62.6
Skip Thoughts	41	29	40	46	52	41.6
InferSent	61	56	68	71	77	66.6
PSIF + PSL	65.7	64.0	74.8	77.3	73.7	71.1
u-SIF + PSL	65.8	65.2	75.9	77.6	72.3	71.4
SCDV + WordVec	64.1	63.9	73.0	76.9	**77.3**	71.0
SCDV + BERT(ctxd)	64.7	64.0	75.4	77.1	73.3	70.9
SCDV + BERT(ctxd) + Anisotropy	**66.8**	64.1	**77.3**	**78.0**	74.6	**72.2**

4. Concept Matching

Embedding	Accuracy	F1
TF-IDF	53.8	70.0
InferSent	54.0	70.1
BERT(pr)	54.8	70.6
SCDV + Word2Vec	53.7	70.0
SCDV + BERT(ctxd)	57.1	73.8
SCDV + BERT(ctxd) + Anisotropy	**58.9**	**74.6**

5. Takeaways

1. Using Contextual representations (like BERT) for WSD can lead to better document representations.
2. **Partition-based** averaging(SCDV) works better than straight word vector averaging.
3. **Anisotropic** approach for **isotropic reduction** are beneficial for getting better document representation.
4. **Fine tuning** of contextual representation such as BERT not beneficial for **low-resource** setting with fewer labeled data.

References

- Dheeraj Mekala, et al *Contextualized Weak Supervision for Text Classification*, ACL 2020
- Dheeraj Mekala*, Vivek Gupta*, et al *Sparse Composite Document Vectors using Soft Clustering over Distributional Semantics*, EMNLP 2017
- Vivek Gupta, et al *P-SIF: Document Embeddings using Partition Averaging*, AAAI 2020

Paper

Code