Efficient Sentence Embedding via Semantic Subspace Analysis

Bin Wang

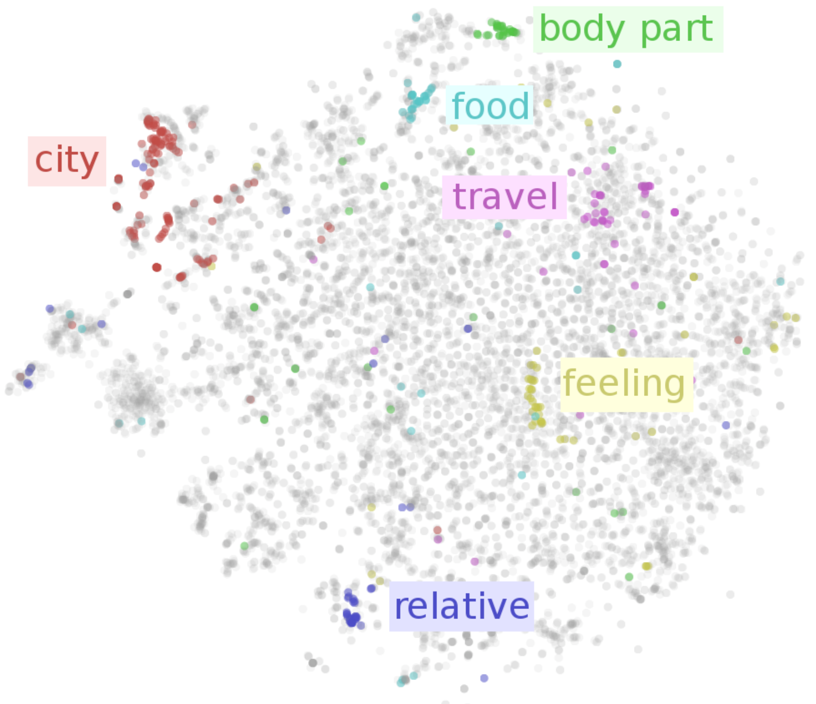
Ph.D. candidate in Electrical Engineering
Viterbi School of Engineering

Background

- Sentence Embeddings:
  - Encode a variable-length input sentence into a fixed size vector
- Examples:
  - Based on Word Embeddings (non-parameterized):
    i. GloVe Averaging
    ii. SIF (Arora et al, 2017)
    iii. Concatenated P-means embeddings (Ruckle'e et al., 2018)
  - Based on RNNs/Transformers (parameterized):
    i. Skip-Thought (Ryan et al., 2015)
    ii. InferSent (Conneau et al., 2017)
    iii. Sentence-BERT (Reimers et al., 2019)
    iv. SBERT-WK (Wang et al., 2020)

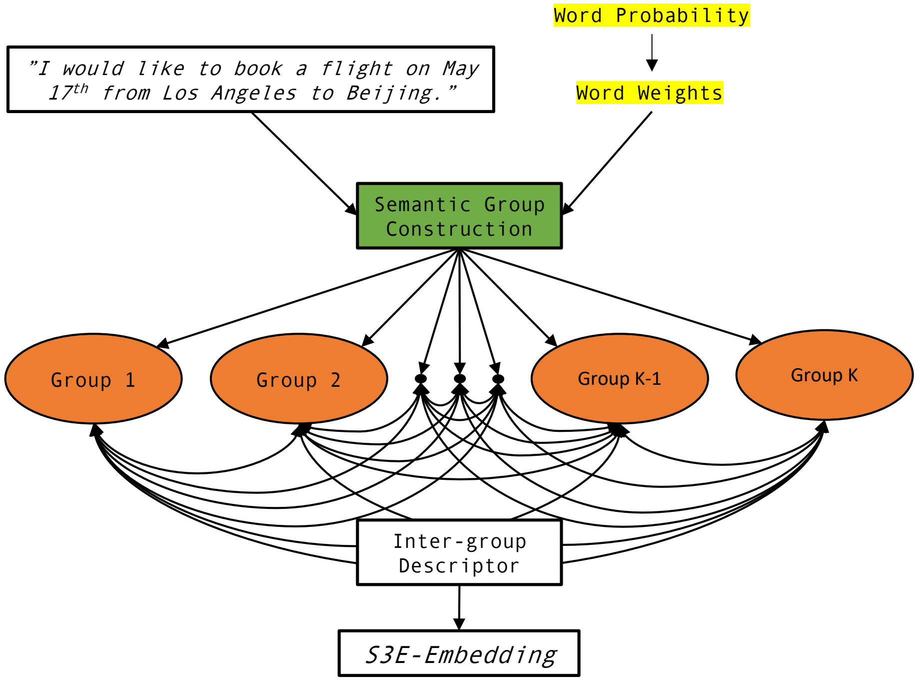
Semantic Grouping Property

- Word embedding naturally forms static groups



Semantic Subspace Sentence Embedding (S3E)

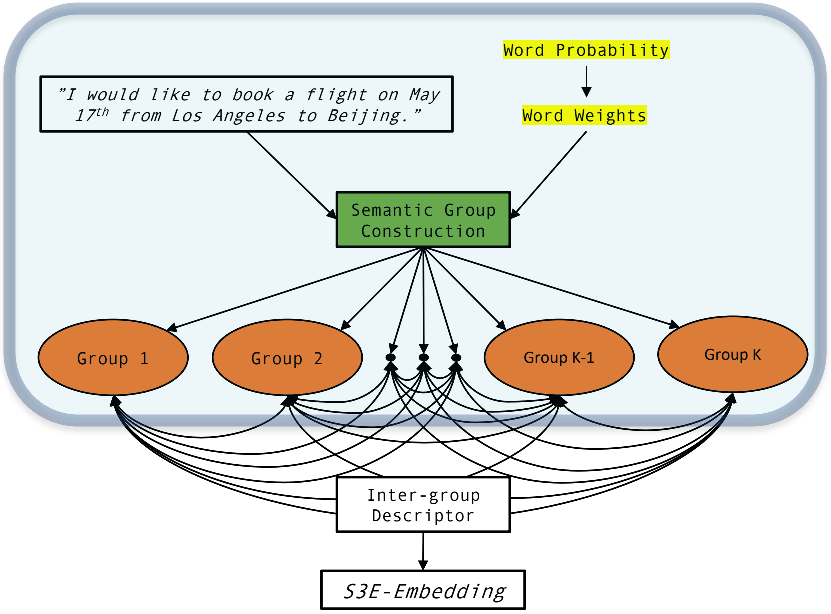
1. Semantic Group Construction
2. Intra-group Descriptor
3. Inter-group Descriptor



Semantic Subspace Sentence Embedding (S3E)

- Semantic Group Construction
  - K-means on word embeddings
  - Word Weights: ‘the’ ‘an’ ‘about’ ‘can’ carries little information
    - High frequency: less weights
    - Low frequency: more weights

$$\text{weight}(w) = \frac{\epsilon}{\epsilon + p(w)}$$



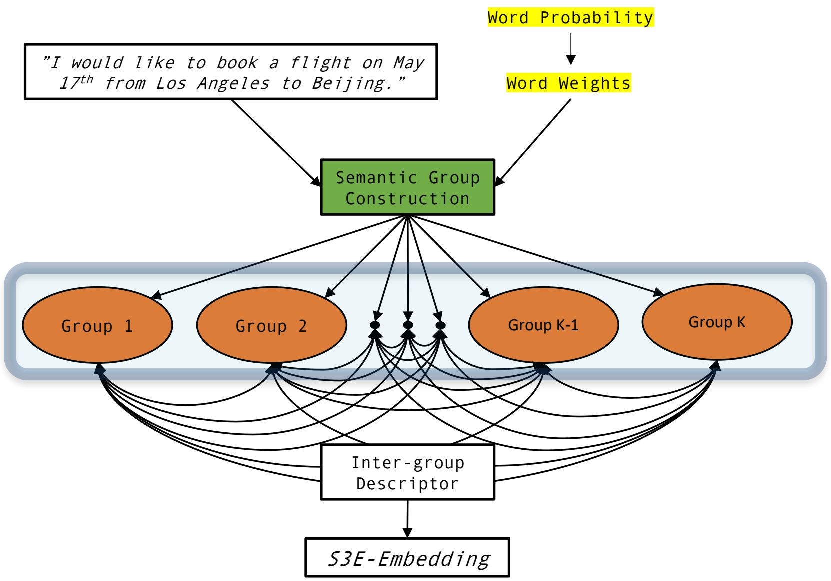
Semantic Subspace Sentence Embedding (S3E)

- Intra-group Descriptor
  1. Centroid Representation

$$g_i = \frac{1}{|G_i|} \sum_{w \in G_i} \text{weight}(w) v_w$$

  2. Residual Representation

$$v_i = \sum_{w \in S \cap G_i} \text{weight}(w)(v_w - g_i)$$

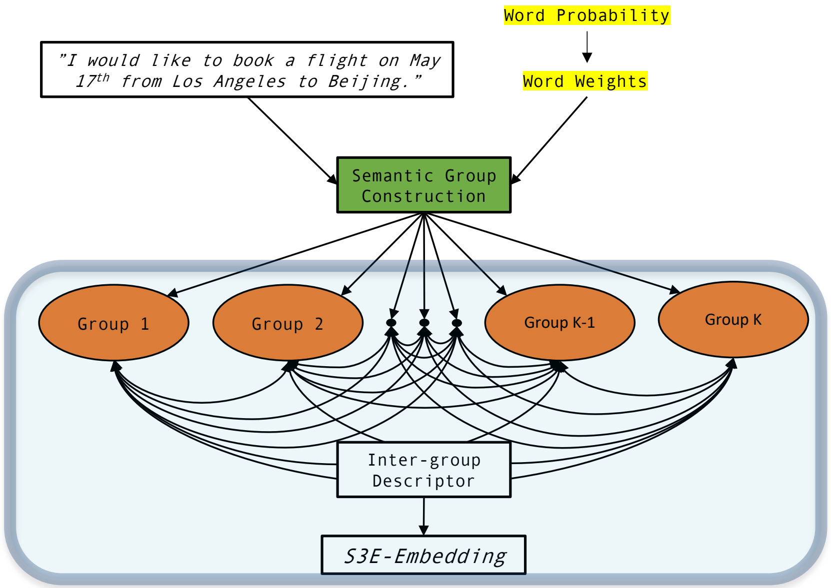


Semantic Subspace Sentence Embedding (S3E)

- Inter-group Descriptor
  - Interaction between semantic groups

$$\Phi(S) = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_K^T \end{bmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{1d} \\ v_{21} & \cdots & v_{2d} \\ \vdots & \ddots & \vdots \\ v_{K1} & \cdots & v_{Kd} \end{pmatrix}_{K \times d}$$

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K} & \sigma_{2K} & \cdots & \sigma_K^2 \end{pmatrix}$$



Experimental Results

- Textual Similarity Tasks

Model	Dim	STS12	STS13	STS14	STS15	STS16	STSB	SICK-R	Avg.
Parameterized models									
skip-thought (Kiros et al., 2015)	4800	30.8	24.8	31.4	31.0	-	-	86.0	40.80
InferSent (Conneau et al., 2017)	4096	58.6	51.5	67.8	68.3	70.4	74.7	**88.3**	68.51
ELMo (Peters et al., 2018)	3072	55.0	51.0	63.0	69.0	64.0	65.0	84.0	64.43
Avg. BERT (Devlin et al., 2018)	768	46.9	52.8	57.2	63.5	64.5	65.2	80.5	61.51
SBERT-WK (Wang et al., 2019)	768	**70.2**	**68.1**	**75.5**	**76.9**	**74.5**	**80.0**	87.4	**76.09**
Non-parameterized models									
Avg. GloVe	300	52.3	50.5	55.2	56.7	54.9	65.8	80.0	59.34
SIF (Arora et al., 2017)	300	56.2	56.6	68.5	71.7	-	72.0	**86.0**	68.50
p-mean (Rucklle et al., 2018)	3600	54.0	52.0	63.0	66.0	67.0	72.0	**86.0**	65.71
S3E (GloVe)	355-1575	59.5	62.4	68.5	72.3	70.9	75.5	82.7	69.59
S3E (FastText)	355-1575	**62.5**	67.8	70.2	**76.1**	74.3	77.5	84.7	72.64
S3E (L.F.P.)	955-2175	61.0	**69.3**	**73.2**	**76.1**	**74.4**	**78.6**	84.7	**73.90**

  - Competitive among non-parameterized models
  - High quality word embedding provides better results

Experimental Results

- Supervised Tasks

Model	Dim	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-E	Avg.
`Parameterized models`										
skip-thought [5]	4800	76.6	81.0	93.3	87.1	81.8	91.0	73.2	84.3	83.54
FastSent [22]	300	70.8	78.4	88.7	80.6	-	76.8	72.2	-	77.92
InferSent [6]	4096	79.3	85.5	92.3	90.0	83.2	87.6	75.5	85.1	84.81
Sent2Vec [21]	700	75.8	80.3	91.1	85.9	-	86.4	72.5	-	82.00
USE [7]	512	80.2	86.0	93.7	87.0	86.1	**93.8**	72.3	83.3	85.30
ELMo [23]	3072	80.9	84.0	94.6	91.0	86.7	93.6	72.9	82.4	85.76
SBERT-WK [10]	768	**83.0**	**89.1**	**95.2**	**90.6**	**89.2**	93.2	**77.4**	**85.5**	**87.90**
`Non-parameterized models`										
GloVe(Ave)	300	77.6	78.5	91.5	87.9	79.8	83.6	72.1	79.0	81.25
SIF [11]	300	77.3	78.6	90.5	87.0	82.2	78.0	-	**84.6**	82.60
p-mean [14]	3600	78.3	80.8	92.6	89.1	**84.0**	88.4	73.2	83.5	83.74
DCT [15]	300-1800	78.5	80.1	92.8	88.4	83.7	**89.8**	75.0	80.6	83.61
VLAWE [18]	3000	77.7	79.2	91.7	88.1	80.8	87.0	72.8	81.2	82.31
S3E (GloVe)	355-1575	78.3	80.4	92.5	**89.4**	82.0	88.2	74.9	82.0	83.46
S3E (FastText)	355-1575	78.8	**81.4**	92.9	88.5	83.5	87.0	**75.7**	81.4	83.65
S3E(L.F.P.)	955-2175	**79.4**	**81.4**	**92.9**	**89.4**	83.5	89.0	75.6	82.6	**84.23**

  - Competitive among non-parameterized models
  - High quality word embedding provides better results

- Inference Time

Model	CPU inference time (ms)
InferSent	53.07
SBERT-WK	179.27
GEM	26.54
SIF	1.56
Proposed S3E	**0.69**

  - Clustering can be pre-computed
  - Low complexity with CPU
  - Mobile device applications

Conclusion & Future Work

- S3E: Non-parameterized model based on word embedding
  - Employ semantic grouping property of word embedding
  - Effective and efficient
  - Modularized design:
    - Exploration on clustering: subspace clustering
    - Exploration on correlation computation: non-linear kernel functions