Regularized and Retrofitted models for Learning Sentence Representation with Context

Tanay Kumar Saha[1] **Shafiq Joty**[2] Naeemul Hassan[3]
Mohammad Al Hasan[1]

[1]Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA

[2]Nanyang Technological University, Singapore

[3]University of Mississippi, Oxford, Mississippi

November 7, 2017

Outline

Outline

Distributed Representation of Sentences

- Represent sentences with **condensed real-valued vectors** that capture syntactic and semantic properties of the sentence
  - *I play soccer* $\Rightarrow$ [0.2, 0.3, 0.4]
- Many sentence-level text processing tasks rely on representing sentences with fixed-length vectors
- The most common approach uses bag-of-ngrams (e.g., tf.idf)
- Distributed representation has been shown to perform better

Motivation

- Most existing Sen2Vec methods disregard **context** of a sentence
- Meaning of one sentence depends on the meaning of its neighbors
  - *And I was wondering about the GD LEV*
  - <u>*Is it reusable?*</u>
  - *Or is it discarded to burn up on return to LEO?*
- Our approach: incorporate **extra-sentential context** into Sen2Vec
- We propose two methods: **regularization** and **retrofitting**
- We experiment with two types of context: **discourse** and **similarity**.

Outline

Our Approach

- Consider **content** as well as **context** of a sentence
- Treat the context sentences as **atomic** linguistic units
  - Similar in spirit to (Le & Mikolov, 2014)
  - Efficient to train compared to **compositional** methods like encoder-decoder models (e.g., SDAE, Skip-Thought)

Content Model (Sen2Vec)

- Treats sentences and words similarly
- Represented by vectors in shared embedding matrix
- **v**: he works in woodworking

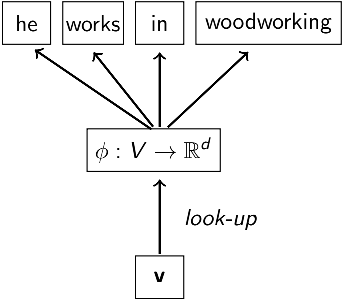


Figure: Distributed bag of words or DBOW (Le & Mikolov, 2014)

Context Types

- Discourse Context
  - Formed by **previous** and **following** sentences in the text
  - Adjacent sentences in a text are logically connected by certain coherence relations (e.g., elaboration, contrast)
- Similarity Context
  - Based on more **direct measures** of similarity (e.g., cosine)
  - Considers similarity with all other sentences
- Context can be represented by a **graph neighborhood**, $\mathcal{N}(\mathbf{v})$

Similarity Network Construction

- Represent the sentences with vectors learned from Sen2Vec, then measure the cosine similarity between the vectors
- Restrict context size of a sentence for computational efficiency
- Set thresholds for intra- and across-document connections
- Allow up to 20 most similar neighbors.

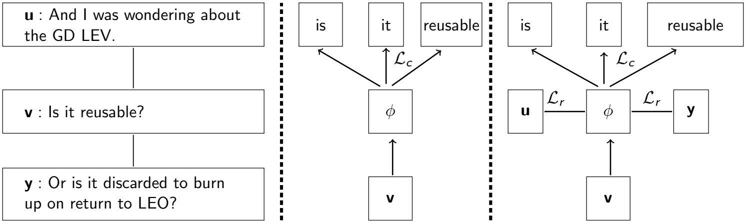
Regularized Models (Reg-dis, Reg-sim)

- Incorporate neighborhood **directly** into the objective function of the content-based model (Sen2Vec) as a regularizer
- Objective function:

$$
\begin{aligned}
J(\phi) &= \sum_{\mathbf{v} \in V} \Big[\mathcal{L}_c(\mathbf{v}) + \beta \mathcal{L}_r(\mathbf{v}, N(\mathbf{v})) \Big] \\
&= \sum_{\mathbf{v} \in V} \Big[\underbrace{\mathcal{L}_c(\mathbf{v})}_{\text{Content loss}} + \underbrace{\beta \sum_{(\mathbf{v},\mathbf{u}) \in E} ||\phi(\mathbf{u}) - \phi(\mathbf{v})||^2 \Big]}_{\text{Graph smoothing}}
\end{aligned} \tag{1}
$$

- Train with SGD
- Regularization with **discourse** context $\Rightarrow$ Reg-dis
- Regularization with **similarity** context $\Rightarrow$ Reg-sim

Pictorial Depiction



(a) A sequence of sentences **(b) Sen2Vec (DBOW)** **(c)** Reg-dis

Retrofitted Model (Ret-dis, Ret-sim)

- **Retrofit** vectors learned from Sen2Vec s.t. the revised vector $\phi(\mathbf{v})$:
  - Similar to the prior vector, $\phi'(\mathbf{v})$
  - Similar to the vectors of its neighboring sentences, $\phi(\mathbf{u})$
- Objective function:

$$J(\phi) = \sum_{\mathbf{v} \in V} \underbrace{\alpha_v ||\phi(\mathbf{v}) - \phi'(\mathbf{v})||^2}_{\text{close to prior}} + \underbrace{\sum_{(\mathbf{v},\mathbf{u}) \in E} \beta_{u,v} ||\phi(\mathbf{u}) - \phi(\mathbf{v})||^2}_{\text{graph smoothing}} \qquad (2)$$

- Solve using **Jacobi iterative** method
- Retrofit with **discourse** context $\Rightarrow$ Ret-dis
- Retrofit with **similarity** context $\Rightarrow$ Ret-sim

Outline

Evaluation Tasks and Datasets

1. **Extractive summarization (ranking task)**
   - Select the most important sentences to form a summary
   - Use the popular graph-based algorithm LexRank
     - nodes $\Rightarrow$ sentences
     - edges $\Rightarrow$ cosine similarity between vectors (learned by models)
   - Benchmark datasets from **DUC-01** and **DUC-02** for evaluation

Dataset	#Doc.	#Avg. Sen.	#Avg. Sum.
DUC 2001	486	40	2.17
DUC 2002	471	28	2.04

Evaluation Tasks and Datasets

1. **Topic classification and clustering**
   - Use learned vectors to classify or cluster sentences into topics
   - MaxEnt classifier and K-means++ clustering algorithm
   - Text categorization corpora: **Reuters-21578** & **20-Newsgroups**.
     - But, we need sentence-level annotation for evaluation
     - Naive assumption: sentences of a document share the same topic label as the document ⇒ induces lot of noise
     - Our approach: LexRank to select top 20% sentences of each document as representatives of the document

Dataset	#Doc.	Total #sen.	Annot. #sen	Train #sen.	Test #sen.	#Class
Reuters	9,001	42,192	13,305	7,738	3,618	8
Newsgroups	7,781	95,809	22,374	10,594	9,075	8

Classification and Clustering Performance

	Topic Classification Results						Topic Clustering Results			
	Reuters			*Newsgroups*			*Reuters*		*Newsgroups*	
	F_1	Acc	κ	F_1	Acc	κ	V	AMI	V	AMI
Sen2Vec	83.25	83.91	79.37	79.38	79.47	76.16	42.74	40.00	35.30	34.74
Tf-Idf	(−) 3.51	(−) 2.68	(−) 3.85	(−) 9.95	(−) 9.72	(−) 11.55	(−) 21.34	(−) 20.14	(−) 29.20	(−) 30.60
W2V-avg	(+) 2.06	(+) 1.91	(+) 2.51	(−) 0.42	(−) 0.44	(−) 0.50	(−) 11.96	(−) 10.18	(−) 17.90	(−) 18.50
C-PHRASE	(−) 2.33	(−) 2.01	(−) 2.78	(−) 2.49	(−) 2.38	(−) 2.86	(−) 11.94	(−) 10.80	(−) 1.70	(−) 1.44
FastSent	(−) 0.37	(−) 0.29	(−) 0.41	(−) 12.23	(−) 12.17	(−) 14.21	(−) 15.54	(−) 13.06	(−) 34.40	(−) 34.16
Skip-Thought	(−) 19.13	(−) 15.61	(−) 21.8	(−) 13.79	(−) 13.47	(−) 15.76	(−) 29.94	(−) 28.00	(−) 27.50	(−) 27.04
RET-sim	(+) 0.92	(+) 1.28	(+) 1.65	(+) 2.00	(+) 1.97	(+) 2.27	(+) 3.72	(+) 3.34	(+) 5.22	(+) 5.70
RET-dis	(+) 1.66	(+) 1.79	(+) 2.30	(+) 5.00	(+) 4.91	(+) 5.71	(+) 4.56	(+) 4.12	(+) 6.28	(+) 6.76
REG-sim	(+) **2.53**	(+) **2.53**	(+) **3.28**	(+) 3.31	(+) 3.29	(+) 3.81	(+) 4.76	(+) 4.40	(+) **12.78**	(+) 12.18
REG-dis	(+) 2.52	(+) 2.43	(+) 3.17	(+) **5.41**	(+) **5.34**	(+) **6.20**	(+) **7.40**	(+) **6.82**	(+) 12.54	(+) **12.44**

Table: Performance on topic classification & clustering in comparison to Sen2Vec

Summarization Performance

	DUC'01	DUC'02
Sen2Vec	43.88	54.01
Tf-Idf	**(+) 4.83**	(+) 1.51
W2V-avg	(−) 0.62	(+) 1.44
C-PHRASE	(+) 2.52	(+) 1.68
FastSent	(−) 4.15	(−) 7.53
Skip-Thought	(+) 0.88	(−) 2.65
RET-sim	(−) 0.62	(+) 0.42
RET-dis	(+) 0.45	(−) 0.37
REG-sim	(+) 2.90	(+) **2.02**
REG-dis	(−) 1.92	(−) 8.77

Table: ROUGE-1 scores on DUC datasets in comparison to Sen2Vec

Outline

Conclusion and Future Work

- Novel models for learning vector representation of sentences that consider not only content of a sentence but also its context
- Two ways to incorporate context: retrofitting and regularizing
- Two types of context: discourse and similarity
- Discourse context beneficial for topic classification and clustering, whereas the similarity context beneficial for summarization
- Explore further how our models perform compared to existing compositional models, where documents with sentence-level sentiment annotation exists

Thanks!

- Code and Datasets: https://github.com/tksaha/con-s2v/tree/jointlearning
- Check our CON-S2V ECML-2017 paper