Domain Adapted Word Embeddings for Improved Sentiment Classification

Prathusha K Sarma[1], Yingyu Liang[2] and William A Sethares[1]

[1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, USA

[2]Department of Computer Sciences, University of Wisconsin-Madison, USA

ABSTRACT:

Generic word embeddings are trained on large-scale generic corpora; *Domain Specific* (DS) word embeddings are trained only on data from a domain of interest. This poster outlines a method to combine the breadth of generic embeddings with the specificity of domain specific embeddings. The resulting embeddings, called *Domain Adapted* (DA) word embeddings are formed by aligning corresponding word vectors using Canonical Correlation Analysis (CCA) or the related nonlinear Kernel CCA. Evaluation results on sentiment classification tasks show that the DA embeddings outperform both generic and DS embeddings when used as input features to standard or state-of-the-art sentence encoding algorithms for classification.

Introduction

Motivation:

- Text data from applications yielding small sized data sets see a heavy influence of domain semantics in language use.
- One example of such a data set comes from the Substance User Disorders (SUDs) data set.
- Such data sets are focused around a specific topic of interest. Language use is often guided by the topic being discussed leading to strong and non standard word associations.
- While these data sets are rich in domain semantics, they are limited by availability making it challenging to train large scale neural network based algorithms for analysis.

Proposed Solution:

- Combine domain semantics along with wider generalizations of word meaning to form domain adapted word embeddings.
- Obtain generic embeddings from an algorithm trained on a large generic data corpus and obtain DS embeddings from a count based embedding technique on a data set from target domain for a fixed sized vocabulary.
- Combine both sets of embeddings to form DA embeddings.

Outline of Process

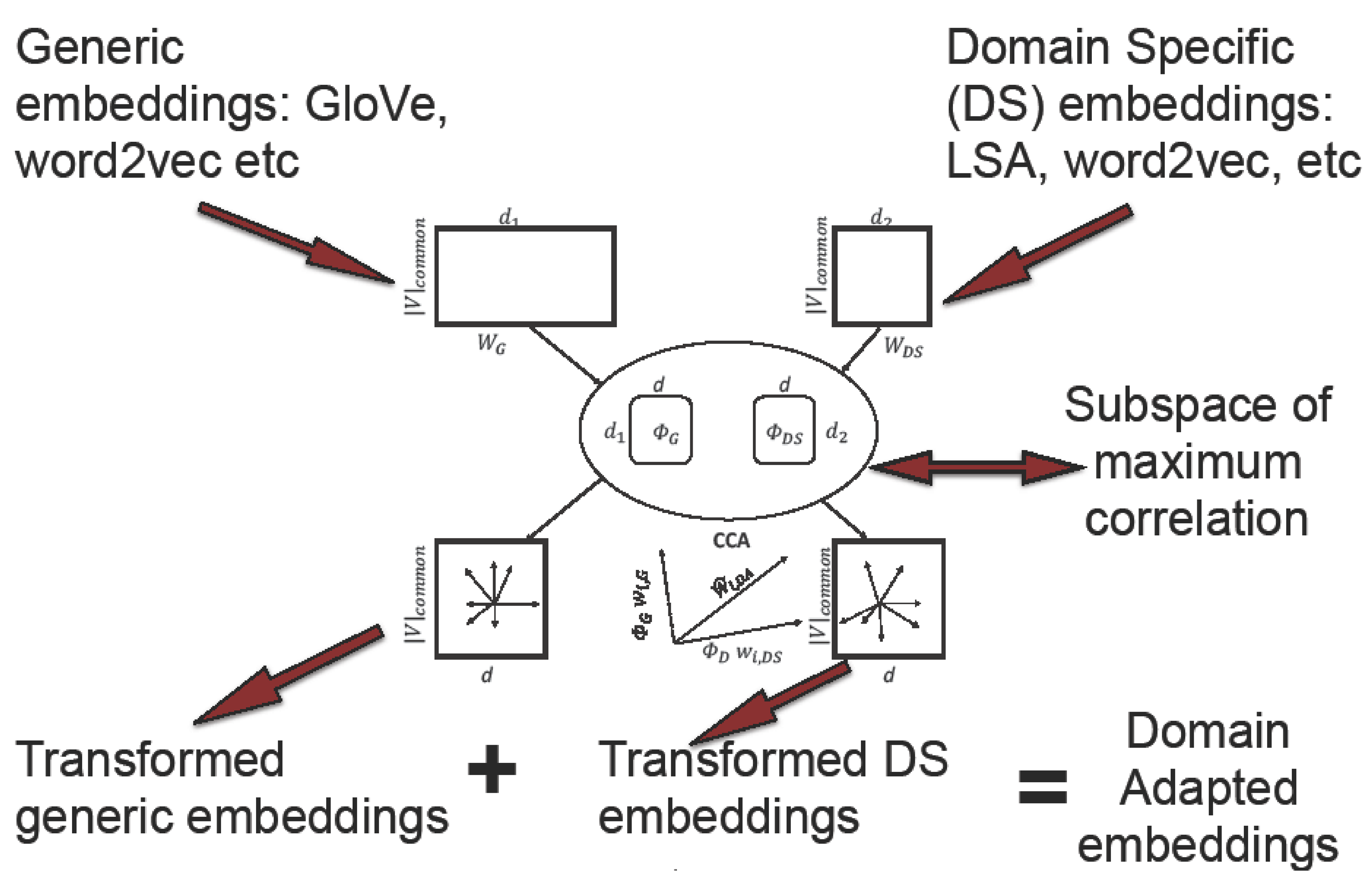

Fig1: This figure illustrates the process of obtaining DA embeddings from Generic and DS embeddings.

CCA/KCCA for Domain Adapted Word Embeddings

- Given vocabulary V, obtain corresponding generic and DS word embeddings $\mathbf{W_G}$ and $\mathbf{W_{DS}}$.
- Generic embeddings such as GloVe are obtained from optimized solvers trained on a generic corpus like Wikipedia.
- DS embeddings can be obtained from count based embedding techniques such as Latent Semantic Analysis (LSA).
- Find a subspace such that correlations between $\mathbf{W_G}$ and $\mathbf{W_{DS}}$ are maximized. Projection directions into desired subspace can be obtained by solving the following optimization,

$$\max_{\Phi_G, \Phi_{DS}} \frac{\mathbb{E}[\langle \mathbf{W}_G, \Phi_G \rangle \langle \mathbf{W}_{DS}, \Phi_{DS} \rangle]}{\sqrt{\mathbb{E}[\langle \mathbf{W}_G, \Phi_G \rangle]^2 \mathbb{E}[\langle \mathbf{W}_{DS}, \Phi_{DS} \rangle]^2}}$$

- Project $\mathbf{W_G}$ and $\mathbf{W_{DS}}$ to obtain,

$$\bar{\mathbf{W}}_G = \mathbf{W}_G \Phi_G$$
$$\bar{\mathbf{W}}_{DS} = \mathbf{W}_{DS} \Phi_{DS}$$

Domain Adapted Word Embeddings: Minimize Sum of Squared Distances

- Express a domain adapted word embedding as,

$$\hat{\mathbf{w}}_{i,DA} = \alpha \bar{\mathbf{w}}_{i,DS} + \beta \bar{\mathbf{w}}_{i,G}$$

- Solve for alpha and beta by minimizing the sum of squared distances between projected generic and DS embeddings,

$$\min_{\alpha,\beta} \|\bar{\mathbf{w}}_{i,DS} - (\alpha \bar{\mathbf{w}}_{i,DS} + \beta \bar{\mathbf{w}}_{i,G})\|_2^2 + \|\bar{\mathbf{w}}_{i,G} - (\alpha \bar{\mathbf{w}}_{i,DS} + \beta \bar{\mathbf{w}}_{i,G})\|_2^2$$

- Solution: alpha=beta=0.5.

Domain Adapted Word Embeddings: Minimize Sum of Cluster Variances

- Assumption1 : beta = 1- alpha.
- Assumption2 : Task is binary classification, labels = positive/negative. Express each document as sum of word embeddings.
- Obtain alpha by minimizing sum of variances of positive and negative document clusters.

$$\min_{\alpha \in [0,1]} \frac{1}{k} \sum_{i=1}^{k} ||d_{p_i} - \mu_p||_2^2 + \frac{1}{N-k} \sum_{i=1}^{N-k} ||d_{n_i} - \mu_n||_2^2$$

Experimental Set Up and Results

Data Sets:

- Balanced data sets of 1000 reviews from Yelp, Amazon and IMDB. Reviews labeled 'positive' or 'negative'.
- Unbalanced A-CHESS data set consisting of 2500 text messages from an AA discussion forum. 8% of messages are 'threat', i.e. indicative of relapse risk.

Baseline Word Embeddings:

- Generic embeddings: GloVe-Wiki (Glv), GloVe -Common Crawl (GlvCC), word2vec.
- Domain Specific embeddings: Latent Semantic Analysis (LSA), re-trained word2vec (DSw2v).
- concSVD: concatenation based baselines. Generic and DS embeddings are concatenated and SVD is performed to obtain DA embeddings.

Baseline Algorithms:

- Standard classification: Text is expressed as weighted sum of constituent word embeddings.
- InferSent: Bidirectional LSTM with max pooling, sentence encoder. This encoder is initialized with generic and DA embedding and performance is compared.
- RNTN: Recursive Neural Tensor Network is a neural network based dependency parser for sentiment classification.

Data Set		Embedding	Avg Precision	Avg F-score	Avg AUC
A-CHESS	DA	**KCCA(Glv, LSA)**	**32.07 ± 1.3**	**39.32 ± 2.5**	**65.96 ± 1.3**
		CCA(Glv,LSA)	32.70 ± 1.5	35.48 ± 4.2	62.15 ± 2.9
		KCCA(w2v, LSA)	**33.45 ± 1.3**	**39.81 ± 1.0**	**65.92 ± 0.6**
		CCA(w2v,LSA)	33.06 ± 3.2	34.02 ± 1.1	60.91 ± 0.9
		KCCA(GlvCC,LSA)	36.38 ± 1.2	34.71 ± 4.8	61.36 ± 2.6
		CCA(GlvCC,LSA)	32.11 ± 2.9	36.85 ± 4.4	62.99 ± 3.1
		KCCA(w2v, DSw2v)	25.59 ± 1.2	28.27 ± 3.1	57.25 ± 1.7
		CCA(w2v, DSw2v)	24.88 ± 1.4	29.17 ± 3.1	57.76 ± 2.0
		concSVD(Glv, LSA)	27.27 ± 2.9	34.45 ± 3.0	61.59 ± 2.3
		concSVD(w2v, LSA)	29.84 ± 2.3	36.32 ± 3.3	62.94 ± 1.1
		concSVD(GlvCC,LSA)	28.09 ± 1.9	35.06 ± 1.4	62.13 ± 2.6
	G	GloVe	30.82 ± 2.0	33.67 ± 3.4	60.80 ± 2.3
		GloVeCC	**38.13 ± 0.8**	**27.45 ± 3.1**	**57.49 ± 1.2**
		word2vec	32.67 ± 3.5	31.72 ± 1.6	59.64 ± 0.5
	DS	LSA	27.42 ± 1.6	34.38 ± 2.3	61.56 ± 1.9
		word2vec	24.48 ± 0.8	27.97 ± 3.7	57.08 ± 2.5

Data Set		Embedding	Avg Precision	Avg F-score	Avg AUC
Yelp	DA	KCCA(Glv, LSA)	85.36 ± 2.8	81.89 ± 2.8	82.57 ± 1.3
		CCA(Glv,LSA)	83.69 ± 4.7	79.48 ± 4.7	80.33 ± 2.9
		KCCA(w2v, LSA)	87.45 ± 1.2	83.36 ± 1.2	84.10 ± 0.9
		CCA(w2v,LSA)	84.52 ± 2.3	80.02 ± 2.6	81.04 ± 2.1
		KCCA(GlvCC,LSA)	**88.11 ± 3.0**	**85.35 ± 2.7**	**85.80 ± 2.4**
		CCA(GlvCC,LSA)	83.69 ± 3.5	78.99 ± 4.2	80.03 ± 3.7
		KCCA(w2v, DSw2v)	78.09 ± 1.7	76.04 ± 1.7	76.66 ± 1.5
		CCA(w2v, DSw2v)	86.22 ± 3.5	84.35 ± 2.4	84.65 ± 2.2
		concSVD(Glv, LSA)	80.14 ± 2.6	78.50 ± 3.0	78.92 ± 2.7
		concSVD(w2v, LSA)	85.11 ± 2.3	83.51 ± 2.2	83.80 ± 2.0
		concSVD(GlvCC,LSA)	84.20 ± 3.7	80.39 ± 3.7	80.83 ± 3.9
	G	GloVe	77.13 ± 4.2	72.32 ± 7.9	74.17 ± 5.0
		GloVeCC	82.10 ± 3.5	76.74 ± 3.4	78.17 ± 2.7
		word2vec	82.80 ± 3.5	78.28 ± 3.5	79.35 ± 3.1
	DS	LSA	75.36 ± 5.4	71.17 ± 4.3	72.57 ± 4.3
		word2vec	73.08 ± 2.2	70.97 ± 2.4	71.76 ± 2.1

Fig 2: Tables show Average Precision, F-score and AUC from standard classification task on unbalanced A-CHESS and balanced Yelp data set. KCCA and CCA DA embeddings are compared against baseline word embeddings. Best performing word embeddings are highlighted in yellow.

Data Set	Embedding	Avg Precision	Avg F-score	Avg AUC
Yelp	GlvCC	86.47 ± 1.9	83.51 ± 2.6	83.83 ± 2.2
	KCCA(GlvCC,LSA)	**91.06 ± 0.8**	**88.66 ± 2.4**	**88.76 ± 2.4**
	CCA(GlvCC, LSA)	86.26 ± 1.4	82.61 ± 1.1	83.99 ± 0.8
	concSVD(GlvCC, LSA)	85.53 ± 2.1	84.90 ± 1.7	84.96 ± 1.5
	RNTN	83.11 ± 1.1	-	-
Amazon	GlvCC	87.93 ± 2.7	82.41 ± 3.3	83.24 ± 2.8
	KCCA(GlvCC,LSA)	**90.56 ± 2.1**	**86.52 ± 2.0**	**86.74 ± 1.9**
	CCA(GlvCC, LSA)	87.12 ± 2.6	83.18 ± 2.2	83.78 ± 2.1
	concSVD(GlvCC, LSA)	85.73 ± 1.9	85.19 ± 2.4	85.17 ± 2.6
	RNTN	82.84 ± 0.6	-	-
A-CHESS	**GlvCC**	**52.21 ± 5.1**	**55.26 ± 5.6**	**74.28 ± 3.6**
	KCCA(GlvCC,LSA)	**55.37 ± 5.5**	**50.67 ± 5.0**	**69.89 ± 3.1**
	CCA(GlvCC, LSA)	54.34 ± 3.6	48.76 ± 2.9	68.78 ± 2.4
	concSVD(GlvCC, LSA)	40.41 ± 4.2	44.75 ± 5.2	68.13 ± 3.8
	RNTN	-	-	-

Fig 3: Table shows results obtained from InferSent encoder initialized with generic and DA embeddings followed by classification. Best performing initialization highlighted in yellow.

Data Set	DA Embedding	α	Avg Precision	Avg F-score	Avg AUC
A-CHESS	**KCCA(Glv, LSA)**	**0.4**	**37.32 ± 1.6**	**41.64 ± 2.8**	**66.13 ± 2.1**
	KCCA(Glv, LSA)	0.5	32.07 ± 1.3	39.32 ± 2.5	65.96 ± 1.3
	KCCA(w2v, LSA)	**0.55**	**35.06 ± 0.9**	**43.44 ± 1.4**	**68.60 ± 1.3**
	KCCA(w2v, LSA)	0.5	33.45 ± 1.3	39.81 ± 1.0	65.92 ± 0.6
	KCCA(GlvCC, LSA)	**0.75**	**38.65 ± 3.1**	**43.02 ± 2.2**	**67.26 ± 2.2**
	KCCA(GlvCC, LSA)	0.5	36.38 ± 1.2	34.71 ± 4.8	61.36 ± 2.6
	GloVe-CC	-	38.13 ± 0.8	27.45 ± 3.1	57.49 ± 1.2
Yelp	KCCA(Glv, LSA)	0.25	84.75 ± 2.2	80.02 ± 2.5	81.13 ± 2.0
	KCCA(Glv, LSA)	**0.5**	**85.36 ± 2.8**	**81.89 ± 2.8**	**82.57 ± 1.3**
	KCCA(w2v, LSA)	**0.45**	**87.74 ± 2.2**	**83.57 ± 2.6**	**84.27 ± 2.4**
	KCCA(w2v, LSA)	0.5	87.45 ± 1.2	83.36 ± 1.2	84.10 ± 0.9
	KCCA(GlvCC, LSA)	**0.6**	**88.84 ± 2.3**	**85.36 ± 2.3**	**85.93 ± 2.0**
	KCCA(GlvCC, LSA)	0.5	88.11 ± 3.0	85.35 ± 2.7	85.80 ± 2.4

Fig 4: Table shows results using KCCA embeddings in the standard classification set up. Value of alpha used to obtain DA embeddings is determined by minimizing sum of cluster variances.