Supplemental Material for Domain Adapted Word Embeddings for Improved Sentiment Classification.

Prathusha K Sarma, Yingyu Liang and William A Sethares

University of Wisconsin-Madison
{kameswarasar,sethares}@wisc.edu,
yliang@cs.wisc.edu

Abstract

Supplemental material provides details about word tokens, embedding dimensions and hyperparameter details.

1 Dimensions of CCA and KCCA projections.

Using both KCCA and CCA, generic embeddings and DS embeddings are projected onto their d largest correlated dimensions. By construction, $d \leq \min(d_1, d_2)$. The best d for each data set is obtained via 10 fold cross validation on the sentiment classification task. Table 2 provides dimensions of all word embeddings considered. Note that for LSA and DA, average word embedding dimension across all four data sets are reported. Generic word embeddings such as GloVe and word2vec are of fixed dimensions across all four data sets.

2 Kernel parameter estimation.

Parameter σ of the Gaussian kernel used in KCCA is obtained empirically from the data. The median (μ) of pairwise distances between data points mapped by the kernel function is used to determine σ. Typically $\sigma = \mu$ or $\sigma = 2\mu$. In this section both values are considered for σ and results with the best performing σ are reported.

3 Word tokens and word embeddings dimensions

Data Set	Word Tokens
Yelp	2049
Amazon	1865
IMDB	3075
A-CHESS	3400

Table 1: This table presents the unique tokens present in each of the four data sets considered in the experiments.

Word embedding	Dimension
GloVe	100
word2vec	300
LSA	70
CCA-DA	68
KCCA-DA	68
GloVe common crawl	300
AdaptGloVe	300

Table 2: This table presents the average dimensions of LSA, generic and DA word embeddings.