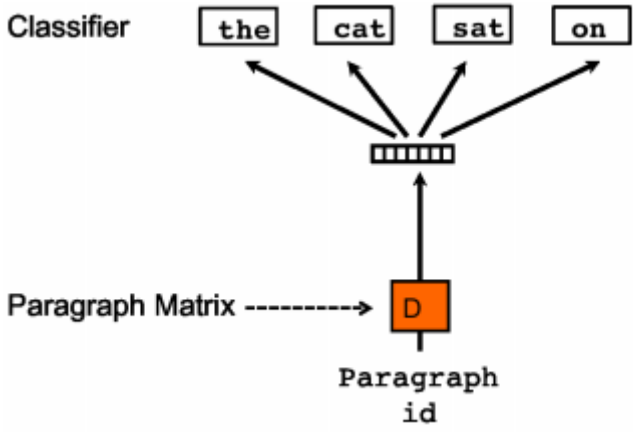
doc2vec的单模型训练

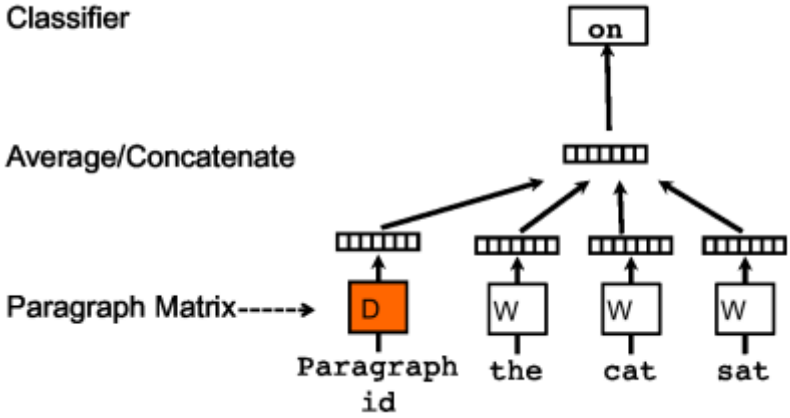
1、模型

1.1 doc2vec model

DBOW



DM



PV-DM w/concatenation - window=5 (both sides) approximates paper's 10-word total window size

将模型输入向量拼接

Doc2Vec(dm=1, dm_concat=1, size=20, window=3, negative=10, hs=0, min_count=1, workers=cores, sample=0),

PV-DBOW

Doc2Vec(dm=0, size=20, negative=10, hs=0, min_count=1, workers=cores, sample=0),

PV-DM w/average 将模型输入向量加和平均

Doc2Vec(dm=1, dm_mean=1, size=20, window=3, negative=10, hs=0, min_count=1, workers=cores, sample=0),

size : doc2vec及词向量的维数，二者同步变化

window（3-5）is the maximum distance between the predicted word and context words used for prediction within a document.文献中是有上下文的。上图没有体现

negative（NEG 随机负采样）2-5 5-20，样本偏少时系数越大越好

sample threshold for configuring which higher-frequency words are randomly downsampled; default is 0 (off), useful value is 1e-5.

1.2 the feature about doc2vec

similarity = cos(vec_answer, vec_standardAnswer)

vec_standardAnswer=mean(all possible standard answer)

衡量相似性的指标：simiarity特征与分数间的相关性，以及该特征对评分性能的提升。

2、数据集介绍

初始训练数据 base
eng_hyt140326数据集 train 977;test 1983; xml 20+20
背景数据：
Kaggle AES手写作文数据 13994 长度小于200的数据 8603（native）
SWECCL2.0： essay,4954; oral, 1426 (nonative)

3、模型训练

3.1 数据的叠加对similarity特征的影响
目的：确定训练文档向量模型的数据集构成。
doc2vec模型：# PV-DM w/concatenation dm=1, dm_concat=1

	parms	pcoef on paper(tst)	pcoef all (dev,tst)	评分性能
baseline				91.52%/0.8169 92.03%/0.8217
base(raw)	size=20; negative=10;w=4	p1:0.65 p2:0.35	0.48 0.49	92.08%/0.8320 92.38%/0.8359
base(raw)	size=50; negative=20;w=4	0.61 0.34	0.46 0.48	91.62%/0.8247 92.08%/0.8298
base(reprocess)	size=50; negative=20;w=4	0.62 0.31	0.45 0.47	91.57%/0.8270 92.08%/0.8330
base+AES_all	size=50; negative=20;w=4	0.69 0.55	0.63 0.63	92.08%/0.8267 92.33%/0.8286
base+AES_all-1426	size=50; negative=20;w=4	0.66 0.52	0.52 0.59	92.13%/0.8273 92.38%/0.8305
base+AES200	size=50; negative=20;w=4	0.66 0.60	0.65 0.63	92.58%/0.8345 92.68%/0.8356
base+AES200-1426	size=50; negative=20;w=4	0.68 0.49	0.54 0.59	92.23%/0.8298 92.28%/0.8318
base+SWE-essay	size=50; negative=20;w=4	0.55 0.39	0.54 0.45	92.73%/0.8330 93.04%/0.8356
base+SWE-essay1426	size=50; negative=20;w=4	0.60 0.42	0.51 0.52	92.18%/0.8315 92.79%/0.8322
base+SWE-oral(1426)	size=50; negative=20;w=4	0.65 0.49	0.50 0.54	92.18%/0.8274 92.84%/0.8300
base+AES200+SWE2	size=50; negative=20;w=4	0.66 0.49	0.64 0.58	92.23%/0.8310 92.63%/0.8356
base+AES200+SWEessay	size=50; negative=20;w=4	0.62 0.43	0.62 0.53	92.48%/0.8314 92.53%/0.8346

小结：

- 总体上加入AES SWE-essay SWE-oral数据集对基于文本相似性的特征提升明显，特别是在特征与分数间的皮尔森相关性上。
- 在训练集中加入相同规模的其它数据集，如base+AES_all-1426、base+AES200-1426、base+SWE-essay1426、base+SWE-oral(1426)这几组实验的评分性能相对baseline都有明显提升；同时，加入语料的文本长度、母语和非母语都会对性能有影响，非母语好于母语的语料（base+SWE-essay1426与base+AES200-1426、base+AES_all-1426），文本长度限制好于不限制（base+AES_all-1426与base+AES200-1426），是否口语语料对评分性能的影响不大（base+SWE-essay1426、base+SWE-oral(1426)）
- 数据的累加对模型在某个试卷上的相似性特征影响是负相关的。
- 后续实验确定的累加数据可以定为base+AES200

3.2 不同模型下的最优模型参数
数据集：base+AES200

	parms	pcoef on paper(tst)	pcoef all (dev,tst)	评分性能
baseline				91.52%/0.8169 92.03%/0.8217
PV-DM c	size=50; negative=20;w=4	p1 0.66 ;p2 0.60	0.65 0.63	92.58%/0.8345 92.68%/0.8356
PV-DM m	size=25; negative=20;w=4	0.44 0.38	0.47 0.41	91.73%/0.8183 92.13%/0.8227
PV-DBOW	size=50; negative=10	0.63 0.27	0.50 0.49	91.83%/0.8180 92.13%/0.8224

3.3 模型拼接

方案：是直接将两种输出的文档向量拼接；PV-DMc+PV-DBOW

	parms	pcoef on paper(tst)	pcoef all (dev,tst)	评分性能
baseline				91.52%/0.8169 92.03%/0.8217
PV-DM c PV-DBOW	DM size=50; negative=20;w=4 DBOW size=50; negative=10	p1 0.73 ;p2 0.41	0.53 0.58	91.47%/0.8215 91.88%/0.8260
PV-DM c/ PV-DBOW				92.33%/0.8311 92.63%/0.8346

PV-DM c/ PV-DBOW：表示分别得到二者的similarity特征，都加入svr训练