4.2 相似度量

由于对于每个单词来说，我们的词嵌入包含了多个向量和不确定参数，所以我们用以下的度量来阐述词与词之间的相似性分数。这些度量选出了与最大相似度相匹配的那部分，也因此指出了那些最相关的意义。

4.2.1 期望可能性内核（Expected LikeLihood Kernel）

对于相似性打分来说，一个自然的选择就是期望可能性内核——不同分布的内积，3.4部分中给出。该度量结合了来自协方差矩阵的不确定性和均值向量间的相似性。

4.2.2 最大余弦相似度（Maximum Cosine Similarity）

这个度量衡量了分布 *f* 和 *g* 中所有成对混合分量的均值向量的相似性。也就是 $d(f,g) = \max_{i,j=1,\ldots,K} \frac{\langle \mu_{f,i}, \mu_{g,j} \rangle}{\|\mu_{f,i}\| \cdot \|\mu_{g,j}\|}$,这个和 *f* 与 *g* 最相似部分的匹配相一致。对于高斯嵌入，最大相似度减小到一般余弦相似度（the usual cosine similarity）。

Word	Co.	Nearest Neighbors
rock	0	basalt:1, boulder:1, boulders:0, stalagmites:0, stalactites:0, rocks:1, sand:0, quartzite:1, bedrock:0
rock	1	rock/:1, ska:0, funk:1, pop-rock:1, punk:1, indie-rock:0, band:0, indie:0, pop:1
bank	0	banks:1, mouth:1, river:1, River:0, confluence:0, waterway:1, downstream:1, upstream:0, dammed:0
bank	1	banks:0, banking:1, banker:0, Banks:1, bankas:1, Citibank:1, Interbank:1, Bankers:0, transactions:1
Apple	0	Strawberry:0, Tomato:1, Raspberry:1, Blackberry:1, Apples:0, Pineapple:1, Grape:1, Lemon:0
Apple	1	Macintosh:1, Mac:1, OS:1, Amiga:0, Compaq:0, Atari:1, PC:1, Windows:0, iMac:0
star	0	stars:0, Quaid:0, starlet:0, Dafoe:0, Stallone:0, Geena:0, Niro:0, Zeta-Jones:1, superstar:0
star	1	stars:1, brightest:0, Milky:0, constellation:1, stellar:0, nebula:1, galactic:1, supernova:1, Ophiuchus:1
cell	0	cellular:0, Nextel:0, 2-line:0, Sprint:0, phones.:1, pda:1, handset:0, handsets:1, pushbuttons:0
cell	1	cytoplasm:0, vesicle:0, cytoplasmic:1, macrophages:0, secreted:1, membrane:0, mitotic:0, endocytosis:1
left	0	After:1, back:0, finally:1, eventually:0, broke:0, joined:1, returned:1, after:1, soon:0
left	1	right-hand:0, hand:0, right:0, left-hand:0, lefthand:0, arrow:0, turn:0, righthand:0, Left:0

Word	Nearest Neighbors
rock	band, bands, Rock, indie, Stones, breakbeat, punk, electronica, funk
bank	banks, banking, trader, trading, Bank, capital, Banco, bankers, cash
Apple	Macintosh, Microsoft, Windows, Macs, Lite, Intel, Desktop, WordPerfect, Mac
star	stars, stellar, brightest, Stars, Galaxy, Stardust, eclipsing, stars., Star
cell	cells, DNA, cellular, cytoplasm, membrane, peptide, macrophages, suppressor, vesicles
left	leaving, turned, back, then, After, after, immediately, broke, end

表1：基于高斯分量的均值向量的余弦相似性的最近的相邻部分：高斯混合嵌入（上），高斯嵌入（下）。表达式 w：i 表示单词 w 的i^{th} 混合分量。

4.2.3 最小欧式距离

在评估嵌入时余弦相似性很常用。但是，在等式（3）中我们的训练目标和欧氏距离直接相关，与例如word2vec中的向量点积相反。因此，我们也考虑了欧式度量：

$d(f,g) = \min_{i,j=1,\ldots,K} [\|\mu_{f,i}\| - \|\mu_{g,j}\|]$

4.3 定性评价

在表1中，我们展示了多义词和他们在嵌入空间中的相领词语的例子，证明了我们的词嵌入训练方法可以捕捉到不同的词义。例如，一个单词比如rock既可能表示stone也可能是rock music，它的每个意思应该用不同的高斯分量表示。我们混合两个高斯模型后得到的结果证实了这个假设，这里我们可以看到rock的0^{th} 个分量与（basalt, boulders）有关，而1^{th} 个分量与（indie， funk， hip-hop）有关。同样的，单词bank的0^{th} 分量代表了河岸，1^{th} 分量代表了金融银行。

相反的，在表1（下）中，可以看见对于单一混合分量的高斯嵌入，多义词的最近邻居显著的只与一个意义有关。例如，rock的邻居主要与摇滚乐有关，bank则主要和金融银行有关。这些多义词的可供选择的意义并没有在词嵌入中很好的体现出来。从数值的例子上来看，rock和stone在 Vilnis and McCallum (2014) 中的高斯表示中余弦相似度为0.029，远远低于本篇文章中rock的0^{th} 分量，其与stone的余弦相似度为0.586。

在那些单词仅有一个流行的意义的案例中，那些混合分量会相当接近。例如，stone的一个分量与（stones, stonework, slab）接近，其他分量与（carving, relic, excavated）接近，这些（单词）在意思上仅有细微的区别。总的来说，混合能够给出很多特性，比如重尾（重尾分布？） 和比单一高斯模型能描述的更多的不确定的有趣的单峰特征。

嵌入可视化（Embedding Visualization） 我们提供了一个交互式的可视化作为代码库的一部分：https://github.com/benathi/word2gm#visualization，这允许实时搜索单词$K = 1, 2, 3$ 分量的最近邻居（在embeddings标签中）。我们采用和表1中类似的符号，w:i表示单词 w 的第 i 分量。例如，当$K = 2$ 时，我们搜寻 bank:0，会发现最近的邻居比如 river:1, confluence:0, waterway:1，这些暗示了bank的$0^(th)$ 分量有 river bank 的意思。另一方面，搜寻 bank:1 会得到例如 banking:1, banker:0, ATM:0，表示这一分量与 financial bank 接近。在$K = 1$ 链接的比较中，我们同样有一个单峰（w2g）的可视化。

另外，我们$K = 3$ 混合分量的高斯模型的嵌入链接可以学习三个不同的含义。例如，cell有三个分量，分别和（keypad, digits）、（incarcerated, inmate）和（tissue, antibody）接近，分别表示cellphone, jail cell, biological cell。由于拥有2个以上含义的单词不多，数量上的限制使得在我们的模型中$K = 3$ 和$K = 2$ 并没有多大差别。因此，为了简洁，我们不在更多展示$K = 3$ 的结果。

4.4 单词相似性

我们在几个标准单词相似性数据集上评估了我们的词嵌入，它们的名字是，SimLex(Hill et al., 2014), WS or WordSim-353, WS-S(similarity), WS-R(relatedness)(Finkelstein et al., 2002), MEN(Bruni et al., 2014), MC(Miller and Charles, 1991), RG(Rubenstein and Goodenough, 1965), YP(Yang and Powers, 2006), MTurk(-287, -771)(Radinsky et al., 2011; Halawi et al., 2012), RW(Luong et al., 2013)。每个数据集包含了一系列单词对，这些单词对由人类给他们的相关性或者相似度打分。

我们计算了标签和我们由词嵌入得到的分数之间的相关系数（Spearman, 1904）。相关系数（Spearman correlation）基于排序的相关性度量，它评估了分数对真实标签的描述情况。

表2展示了这些相关性结果，利用了期望可能性内核（expected likehood kernel）、最大余弦相似性（maximum cosine similarity）和最小欧氏距离(maximum Euclidean distance)。

Dataset	sg*	w2g*	w2g/mc	w2g/el	w2g/me	w2gm/mc	w2gm/el	w2gm/me
SL	29.39	**32.23**	29.35	25.44	25.43	29.31	26.02	27.59
WS	59.89	65.49	71.53	61.51	64.04	**73.47**	62.85	66.39
WS-S	69.86	76.15	76.70	70.57	72.3	**76.73**	70.08	73.3
WS-R	53.03	58.96	68.34	54.4	55.43	**71.75**	57.98	60.13
MEN	70.27	71.31	72.58	67.81	65.53	**73.55**	68.5	67.7
MC	63.96	70.41	76.48	72.70	**80.66**	79.08	76.75	80.33
RG	70.01	71	73.30	72.29	72.12	**74.51**	71.55	73.52
YP	39.34	41.5	41.96	38.38	36.41	**45.07**	39.18	38.58
MT-287	-	-	64.79	57.5	58.31	**66.60**	57.24	60.61
MT-771	-	-	**60.86**	55.89	54.12	60.82	57.26	56.43
RW	-	-	28.78	32.34	33.16	28.62	31.64	**35.27**

表2： 单词相似性数据集上的相关系数。模型 sg, w2g, w2gm 表示 word2vec 连续跳跃元语法，高斯嵌入，高斯混合嵌入（$K = 2$）。度量 mc, el, me 分别表示最大余弦相似性， 期望可能性内核，和最小欧氏距离。对于 w2g 和 w2gm ， 我们在有最高分的相似性度量下加了下划线。对每一个数据集，我们对所有模型中的最高分进行了加粗。sg*, w2g* 的相关性分数来自 vilnis and McCallum (2014)并与余弦距离相一致。

我们展示了我们的高斯混合模型的结果，并和 word2vec 以及Vilnis and McCallum(2014)提出的原始高斯嵌入的结果相比较。我们注意到在大多数数据集上，我们的单峰高斯嵌入模型也比原始模型表现更好，后者在模型的超参数和初始化中有些不同。

在很多数据集，WS, WS-R, MEN, MC, RG, YP, MT-287, RW上，我们的多标准模型 w2gm 同样比连续跳跃元语法（skip-gram）和高斯嵌入表现更好。在大多数数据集中最大余弦相似性表现最好，然而在 MC 和 RW 中，最小欧氏距离更好。这些结果在单一标准模型和多标准模型中保持一致。

我们同样和 Huang et al.(2012), Neelakantan et al.(2014)提出的多标准模型在 WordSim-353 数据集上进行了比较，并在表3中展示。我们发现，我们的单一标准模型 w2g 和 Huang et al.(2012) 相比很有竞争力，即使是在没有去掉停用词的语料库上。这可能是因为通过协方差学习产生的自动校准，它可以降低那些超高频单词的重要性，比如 the, to, a 等等。另外，我们的多标准模型在 WordSim-353 数据集上大体上胜过了 Huang et al.(2012)的模型和 Neelakantan et al.(2014)的 MSSG模型。

MODEL	$\rho \times 100$
HUANG	64.2
HUANG*	71.3
MSSG 50D	63.2
MSSG 300D	71.2
W2G	70.9
W2GM	**73.5**

表3：我们的混合高斯词向量嵌入在 WordSim353 上的相关系数(ρ)，以及 Huang et al.(2014) 的多标准模型和 Neelakantan et al.(2014)的 MSSG 模型。Huang* 采用去除所有停用词的数据训练。MSSG 300D 的维度为300， 其他为50，不过前者仍然表现不如我们的 w2gm 模型。

4.5 多义词的单词相似性

我们采用了 Huang et al.(2012)介绍的SCWS数据集，它选择了那些由多义词和同音异义的单词构成的单词对。

我们将我们的方法和 Huang(Huang et al., 2012), Tian(Tian et al., 2014), Chen(Chen et al., 2014)提出的模型以及(Neelakantan et al., 2014)提出的MSSG模型进行比较。我们注意到，Chen 的模型采用了外部词汇圆 WordNet，这使它该模型有着额外的优势。

我们用了很多参数来计算相关系数（Spearman correlation）的分数。Maxsim 指最大余弦相似性，AveSim 指关于分量概率的余弦相似性的平均值。

在表4中，w2g 模型在所有单一标准模型中表现最好，无论词向量维度是50还是200。我们的w2gm与其他多标准模型相比表现的也很有竞争力。在 SCWS 中，转移到概率密度方法的灵活性的增益支配了采用一个多标准模型的表现。在大多数其他案例中，我们发现 w2gm 优于 w2g，这里多标准结构在优异表现上和概率表示一样重要。注意到，其他模型同样采用了 AvgSimC 度量， 它用了上下文信息能够产生更好的相关性(Huang et al., 2012; Chen et al., 2014)。我们报告了用现有模型得到的 AvgSim 或者 MaxSim 的数值， 这些模型在 MaxSim 上比我们的模型更有竞争力。

MODEL	DIMENSION	$\rho \times 100$
WORD2VEC SKIP-GRAM	50	61.7
HUANG-S	50	58.6
W2G	50	**64.7**
CHEN-S	200	64.2
W2G	200	**66.2**
HUANG-M AVGSIM	50	62.8
TIAN-M MAXSIM	50	63.6
W2GM MAXSIM	50	62.7
MSSG AVGSIM	50	**64.2**
CHEN-M AVGSIM	200	**66.2**
W2GM MAXSIM	200	65.5

表4：在 SCWS 数据集上的相关系数 ρ 。我们展示了单一标准模型（上）和多标准模型（下）的结果。后缀 -(S, M)指单一和多标准模型。