哈爾濱工業大學

毕业设计（论文）中期报告

题 目：基于 CCA 的跨语言语义表示的研究与实现

专　　业　计算机科学与技术

学　　生　白雪峰

学　　号　1130310108

指导教师　曹海龙

日　　期　2016.04.25

哈尔滨工业大学教务处制

1．论文工作是否按开题报告预定的内容及进度安排进行

1.1　从开题至今已经完成开题报告中预定的内容：

1. 完成对于原模型(BiCCA)的基本考察实验
2. 针对原模型只能考察线性关系的特点提出基于 KCCA 的新模型(BiKCCA)

1.2　除此之外还完成了以下工作：

1. BiKCCA 模型的理论分析和代码实现
2. BiKCCA 模型与 BiCCA 模型的对比实验

2．已完成的研究工作及成果

2.1　对原模型 BiCCA 的基本考察实验

对原模型的考察主要分为以下两部分：

2.1.1　对 CCA 的进一步考察

CCA 模型主要用于寻找两个 X，Y 空间内的投影变量 w,v,使得 X，Y 分别投影到 w,v 的方向后的 X’,Y’间的相关系数最大，也就是使 X’,Y’的“联系”最大化。

除了 NLP 领域外，CCA 模型在其他领域如：图像处理，心理测试等也有着广泛的应用，并且有着不错的效果，所以有理由相信基于 CCA 的跨语言模型在 NLP 领域会有不俗的表现

2.1.2　BiCCA 模型的相关实验

完成了 BiCCA 模型的代码实现工作，并对 BiCCA 模型生成的词向量进行了英文词相似度，双语字典生成，跨语言文本分类三个实验，实验结果将在后面 CCA 同 KCCA 模型的比较中集中展示。

2.2　BiKCCA 模型的提出

在对 BiCCA 模型的试验中，我们发现 BiCCA 模型的效果不是十分的理想，就双语字典生成任务来看，各个语言对模型的 Pearson 相关系数均已经达到 0.94，理论上已经捕获了几乎所有的线性关系，然而多数双语模型正确率只有 70%左右，有的甚至不足 60%；这让我们怀疑是不是语言中还存在 CCA 所捕捉不到的相当数量的非线性关系。

通过考察相关文献，我们发现在其他领域的工作：

(1) Facial expression recognition using kernel canonical correlation analysis(W zheng et ieee,2006)

(2) Appearance models based on kernel canonical correlation analysis(T Melzer et els.,2003)

(3) Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis(Y Yamanishi et bio.,2003)

都选择了使用 KCCA 模型来捕捉非线性因素进而提升性能，而且获得了可观的提升；因此我们也考虑实现一个基于 KCCA 的跨语言模型,期待借此捕捉语言中的非线性因素。

2.3 BiKCCA 模型的理论分析

BiKCCA 模型相比于原模型的主要改动在于 KCCA 模型，以下主要介绍 KCCA 模型

2.3.1 KCCA 原理

上节中已经提到，KCCA 的提出是为了解决 CCA 的非线性相关问题；

KCCA 的主要思想就是利用核函数将原来的低维向量升维到高维空间，这样原来在低维的非线性相关关系到高维就会变成线性相关关系；我们将这里用到的核函数抽象为：

$\Phi: X = (X1X2 \cdots Xn) \to \Phi(X) = \big(\varphi 1(X), \varphi 2(X) \cdots \varphi N(X)\big)$其中 $\mathrm{n} < \mathrm{N}(1)$

2.3.2 KCCA 求解

类似于 CCA,我们的目的也是构造高维空间投影向量$c, d \in R^N$，使$\mathrm{X}' = c^T X, Y' = d^T Y$间的相关系数最大。然而这样直接去求有如下问题：

• 核函数如高斯径向基核函数会将向量升到无穷维，此时 c,d 自然也是无穷的，自然无法求解

• 引入核函数的一个目的就是利用核函数的核矩阵 K，也就是期望出现 $\mathrm{X}^T X$ 型，替换成核矩阵 K，而上面的做法显然没有$\mathrm{X}^T X$ 型，因此很难利用核矩阵，求解比较复杂

重新定义：$P = \begin{bmatrix} \Phi(X1)^T \\ \vdots \\ \Phi(\mathrm{Xm})^T \end{bmatrix}$,则 $P^T = \big[\Phi(X1)\Phi(X2) \cdots \Phi(Xm)\big]$，同理 $Q = \begin{bmatrix} \Phi(Y1)^T \\ \vdots \\ \Phi(\mathrm{Ym})^T \end{bmatrix}$（这里 X,Y 的样本数量是相同的）

令 $\mathrm{c} = P^T \mathrm{a}$，$\mathrm{d} = \mathrm{Q}^T b$，这里 a 相当于通过控制每个高维向量的权重控制 c。（这里用到了 KPCA 的相关证明，空间中的任一向量（哪怕是基向量）都可以由该空间中的所有样本线性表示）

与之对应的：

$$X' = c^T P, Y' = d^T Q \tag{2}$$

类似于 CCA,我们要最大化的还是：

$$\rho(\mathrm{X'}, \mathrm{Y'}) = \frac{\mathrm{cov}(\mathrm{X'}, \mathrm{Y'})}{\sqrt{(Var(X'))}\sqrt{(Var(Y')}} \tag{3}$$

其中：

$$\mathrm{Var(X')} = \mathrm{c^T var(P) c} = \mathrm{a^T P var(P) P^T a} \tag{4}$$

$$\mathrm{Var(Y')} = \mathrm{d^T var(Q) d} = \mathrm{b^T Q var(Q) Q^T b} \tag{5}$$

$$\mathrm{Cov(X', Y')} = \mathrm{c^T cov(P, Q) d} = \mathrm{a^T P cov(P, Q) Q^T b} \tag{6}$$

进行下一步前，先对 P,Q 进行归零化处理,即∑xi=0，好处就是：

$$\mathrm{cov(P, Q)} = 1/\mathrm{N\text{-}1}\ \mathrm{P^T Q}$$

引入核矩阵 $\mathrm{K}, \mathrm{K_p} = \mathrm{PP^T}, \mathrm{K_q} = \mathrm{QQ^T}$

于是上面的公式也可以写成：

$$\mathrm{Var(X')} = \mathrm{a^T P var(P) P^T a} = \mathrm{a^T PP^T PP^T a} = \mathrm{a^T K_p{}^2 a} \tag{7}$$

$$\mathrm{Var(Y')} = \mathrm{b^T Q var(Q) Q^T b} = \mathrm{b^T QQ^T QQ^T b} = \mathrm{b^T K_q{}^2 b} \tag{8}$$

$$\mathrm{Cov(X', Y')} = \mathrm{a^T P cov(P, Q) Q^T b} = \mathrm{a^T PP^T QQ^T b} = \mathrm{a^T K_p K_q b} \tag{9}$$

因此

$$\rho(\mathrm{X'}, \mathrm{Y'}) = \frac{\mathrm{cov(X', Y')}}{\sqrt{(Var(X'))}\sqrt{(Var(Y')}} = \frac{a^T \mathrm{KpKqb}}{\sqrt{(a^T Kp^2 a)}\sqrt{(b^T Kq^2 b}} \tag{10}$$

与 CCA 类似，此时的优化问题变成：

Maximize: $a^T K_p K_q b$

Subject to: $a^T K_p{}^2 a = 1, b^T K_q{}^2 b = 1$

构造 Lagrangian 等式:

$$L = a^T K_p K_q b - \lambda(a^T K_p{}^2 a - 1)\ /2 - \theta(b^T K_q{}^2 b - 1)/2 \tag{11}$$

类似于 CCA 的求解过程，最后也可以得到一个求矩阵特征值的方程：

$$K_pK_pa - \lambda^2K_pK_pa = 0 \quad (12)$$

即：

$$Ia = \lambda^2 a \quad (13)$$

得 $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1$,发现所有的 λ 都是相等的，显然模型是过拟的，需要正则化处理。

2.3.3 正则化的 KCCA

有两种常见的正则化的方法，一种是对 Lagrange 加入正则项，一种是对限制条件做一下修改。

第一种是引入正则项的 Lagrange 函数为：

$$L = a^TK_pK_qb - \lambda(a^TK_p^2a - 1)/2 - \theta(b^TK_q^2b - 1)/2 + \eta(\|a\|^2 + \|b\|^2)/2 \quad (14)$$

同样令 Lagrange 函数的导数为 0，得：

$$K_pK_qb = (\lambda K_p^2 + \eta I)a \quad (15)$$

$$K_qK_pa = (\theta K_q^2 + \eta I)b \quad (16)$$

最后化简可以求得：

$$(K_p + \eta I)^{-1}K_q(K_q + \eta I)^{-1}K_qa = \lambda^2 a \quad (17)$$

最后只要求特征值和特征向量就可以了。

第二种是对限制条件作修改，修改后的模型为：

$$\text{Maximize: } a^TK_pK_qb$$

$$\text{Subject to: } (1-\sigma)a^TK_p^2a + \sigma a^TK_pa = 1 \ (1-\sigma)b^TK_q^2b + \sigma b^TK_qb = 1$$

继续构造 Lagrange 函数，最后也可以化简为与方案 1 类似的结果，这里就不详细推导了，本模型中实现的是第一种正则项的 Lagrange 函数。

2.4 BiKCCA 模型与 BiCCA 模型的对比实验

2.4.1 实验数据

考虑到不同语系语言的不同特性，我们选择了如下 5 个语言对，分别是英语-法语（en-fr），英语-德语(en-de)，英语-捷克语(en-cs)，英语-匈牙利语(en-hu),英语-汉语(en-zh)；其中捷克语，法语，德语分别属于印欧语系的斯拉夫，罗曼，日耳曼语族，匈牙利语属于乌拉尔语系。汉语属于汉藏语系。

实验中所用到的单语语料均来自 Waleed Ammar 等人发布的 the Leipzig Corpora Collection 和 Europarl 语料的结合版本

2.4.2 模型的训练与参数

为了保证实验公平对比，单语词向量均使用 word2vec 的 skip-gram model with negative sampling (Mikolov et al., 2013a) with window of size 5 (tuned over {5, 10, 20}).词向量的维度是 200 维

训练好单语模型后，我们由平行语料分别对 en-fr,en-de,en-cs,en-hu,en-zh 五个语言对挑选了最常用的 6.9K,7.1K,7.0K,7.3K 词生成了双语词典。

对于双语模型，我们使用 k=0.5(tuned over {0.2, 0.3, 0.5, 1.0}) 作为单语和双语模型向量维度的比例系数，这样就得到了维度是 100 的双语词向量。

对于 KCCA,我们两种语言均采用的是 rbf 核，参数 gama 的范围是{1e-1，1e-2,1e-3,1e-4,1e-5}.

2.4.3 词向量的评估

我们主要通过以下三个任务来评估词向量：

- 英语的单语词相似度实验

  我们使用的是 QVEC-CCA（Waleed Ammar et 2016),QVEC 是 Tsvetkov 等人在 2015 年提出的用于评估英语词向量质量的内在指标，QVEC-CCA 将 QVEC 和 CCA 相结合，通过最大化词向量矩阵 X 和语言本身的“特性”矩阵 S 间的线性关系训练模型，最后通过相关系数来作为衡量词向量质量的指标。

- 跨语言字典生成实验

  对于跨语言字典生成任务，我们参考的是 Vulic and Moens (2013a)的做法，但与它不同的是我们人工生成了一个 gold 字典，使用的语料来源是 Open Multilingual WordNet data released by Bond and Foster (2013)，我们去掉了那些低频的词汇，最后分别为 en-fr,en-de,en-cs,en-hu,en-zh 生成了 1.5K,1.4K,1.7K,1.5K,1.6K 个词的 gold 字典。

- 跨语言文本分类实验

  对于跨语言文本分类任务（CLDC），我们参照的是 Klementiev et al.(2012)的做法。我们将此实验扩展到我们做要用到的语言对中,此实验用到的预料来自于 the RCV2 Reuters multilingual corpus，对于一个语言对（L1,L2）,我们在 L1 的语料上训练文本分类器，在 L2 的语料上进行测试。由于只在一个语言上进行了监督学习，在其他语言并没有进行监督学习，所以 CLDC 可以很好的考察我们的跨语言表示方式在不同语言间是否可以保持语义一致性。

  对于 CLDC 模型，用到的词向量来自于之前$2.4.2 中得到的词向量，我们只利用 RCV2 数据去训练文本分类器，参照 Klementiev et al.(2012)的做法,我们就得到一个基于平均感知机的文本分类器。

前两个任务主要是从相似度的角度测量通过双语训练后的词向量究竟相较于单语可以获得怎样程度的提升。第三个任务用于衡量训练好的向量的跨语言能力，可以极大地方便模型的跨语言语义应用。

2.4.3.1 单语词相似度评估

表 1：英文词向量的单语词相似度评估结果

L1	L2	Mono	BiCCA	BiKCCA
En	Fr	**0.39**	37.7	38.2
	De	**0.39**	37.3	37.7
	Cs	**0.39**	37.6	38.3
	Hu	**0.39**	36.9	37.5
	Zh	0.39	37.1	**40.2**
Avg.		**0.39**	37.3	38.4

经 Tsvetkov 的方法评估后得到的实验结果，分数越高表示结果越好。所有结果中最好的由黑体标注，双语词向量中最好的由下划线标注

表 1 中展示了单语模型，BiCCA 模型，BiKCCA 模型生成的英语词向量的评估结果可以看出 BiKCCA 模型相比于 BiCCA 模型均有小幅提升，但

是提升幅度有限。令人惊喜的是 BiKCCA 模型在中英训练后的词向量相比于 BiCCA 提升了 3 个百分点，说明中英间的非线性因素确实起到了一定作用。

有趣的一点是无论是 BiCCA 还是 BiKCCA,qvec 的得分均不如本身的单语词向量，一个比较合理的解释是 QVEC 是通过为英语制作的语言性质矩阵来衡量单语词向量，而对于 BiCCA,BiKCCA，训练的数据只是双语字典，并不能得到语言本身的一些性质，所以得分自然不会很高。

2.4.3.2 跨语言字典生成

表 2：跨语言字典生成任务评估结果

L1	L2	BiCCA			BiKCCA		
		Top-1	Top-10	Top-50	Top-1	Top-10	Top-50
	Fr	53.4	73.2	79.8	**53.5**	**74.2**	**79.9**
	De	54.8	73.3	80.4	**55.0**	**73.9**	**80.8**
En	Cs	36.9	66.4	75.4	**38.3**	**67.9**	**76.6**
	Hu	34.5	54.8	64.5	**37.4**	**58.8**	**67.0**
	Zh	38.3	62.1	73.1	**46.1**	**68.6**	**76.8**
	Avg.	43.6	66.0	74.6	**46.0**	**68.9**	**76.2**

表中数字表示跨语言字典生成 top-k 的精确度，top-k 表示每个单词的最近的 k 个邻近词，k={1,10,50},每一项中最好的结果由黑体标注。若以 MMR(mean reciprocal rank)为评估标准也会获得类似的结果。

表 2 展示了 BiCCA 和 BiKCCA 两个模型在跨语言字典生成任务上的性能；从表中可以发现，对于语言差距较小的 en-fr,en-de，BiKCCA 相对于 BiCCA 的提升并不明显。对于同属印欧语系，语言差别相对较大的 en-cs,BiKCCA 有接近%1 的提升，而对于语言差距很大的 en-hu,en-zh 语言对，BiKCCA 得到的提升是明显的，这也间接表示这两个语言对中存在同义词间一定量的非线性因素。

2.4.3.3 跨语言文本分类

跨语言文本分类的实验由于缺少实验语料，只得到了英德语料的实验结果：

表 3：跨语言文本分类任务结果

L1-L2	BiCCA			BiKCCA		
	P@500	P@1000	P@5000	P@500	P@1000	P@5000
En-De	87.1	84.7	87.1	**89.6**	**88.0**	**87.5**
De-En	**71.5**	73.5	**71.3**	**71.5**	**73.6**	71.0

表中数字表示跨语言文本分类的精确度，P@K 中的 K 指训练和测试的文本的数量，L1-L2 表示在 L1 上训练，L2 上测试。

从表中可以发现，En-De 实验中，BiKCCA 相比于 BiCCA 取得了比较明显的提升，在 De-En 实验中，两个模型的性能相近，BiKCCA 带来的提升非常有限。

3．后期拟完成的研究工作及进度安排

3.1 后期工作

1. 完成跨语言文本分类实验

2. 打算继续补充实验，找出哪些词间是线性关系，哪些词间是非线性关系，并以图表的形式展示出来
3. 对实验结果的进一步分析
4. 整合系统，整合实验结果

3.2 进度安排

- 第 10—13 周 完成补充实验，找出 BiKCCA 模型区别于 BiCCA 模型所捕获的非线性关系
- 第 14—15 周 实验结果的整理与分析，参与结题答辩

4．存在的问题与困难

1. 非线性关系的捕获与展现

目前并没有确定具体如何去捕获非线性关系的词对，可能要花一段时间去寻找解决方案。

2. 对实验结果的一个科学而完整的分析与解释

由于个人水平问题，对结果的分析工作可能存在不严谨的地方，本人将尽力保证分析工作科学严谨。

5．论文按时完成的可能性

可以保证论文按时完成。