基于 CCA 的跨语言语义表示的研究与实现

白雪峰

院　（系）:　计算机科学与技术学院　专　　业:　计算机科学与技术

学　　号:　1130310108　指导教师:　曹海龙

2016 年 06 月

哈爾濱工業大學

毕业设计（论文）

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　　号　1130310108

学　　生　白雪峰

指 导 教 师　曹海龙

答 辩 日 期　2016 年 06 月

摘　要

多语词向量是对两种或多种语言在一个共有的连续的向量空间的一种跨语言表示方式。近几年来，跨语言词向量被广泛应用在各种各样的 NLP 任务中。一个简单却又有效的生成跨语言词向量的方法就是使用典型相关分析（CCA）。然而，CCA 方法工作的前提是假设不同语言间语义相近的词对应的词向量间的关系是线性关系。这种假设在我们看来并不总是正确的，尤其是那些本身差距比较大的语言之间的关系并不能被线性关系描述完全。为了进一步捕捉语言间的关系，我们提出使用更有效的核典型相关分析（KCCA）技术，既捕捉线性关系，也捕捉非线性关系。我们在三个任务（单语词相似度，跨语言词典生成，跨语言文本分类）、五个语言对上对我们模型生成的词向量做了充分的评估，最后表明我们提出的模型相比于原基于 CCA 技术的模型能生成更高语义质量的词向量，尤其是在本身差异比较大的语言间效果更加显著。

关键词：跨语言词语表示；核典型相关分析（KCCA）；词向量评估

Abstract

Cross-lingual word embeddings are representations for vocabularies of two or more languages in one common continuous vector space and are widely used in various NLP tasks. A simple yet efficient way to generate cross-lingual word embeddings is using canonical correlation analysis (CCA). However, CCA works with the assumption that the vector representations of similar words in different languages are related by a linear relationship. This assumption does not always hold true, especially for substantially different languages. We therefore propose to use kernel canonical correlation analysis (KCCA) to capture non-linear relationships between word embeddings of two languages. By extensively evaluating the resulting word embeddings on three tasks (word similarity, cross-lingual dictionary induction, cross-lingual document classification) across five language pairs, we show that our approach produces essentially better semantic vectors than CCA-based method, especially for substantially different languages.

Keywords: Cross-lingual word representation, kernel canonical correlation analysis (KCCA), word embedding evaluation

目　录

Contents

第 1 章 绪论

1.1 课题背景介绍

不久以前，NLP 中最直观，最常用的对语言中单词的数学表示方法是 One-hot Representation，这种方法将每个词表示为一个很长的向量，这个向量的维度是词表大小，所有维度中只有一个维度的值为 1，其他维度都是 0，数值为 1 的维度就代表了当前的词。这种表述显然是存在问题的，它过于庞大而且不能体现词之间的关系。而近年来提出的词向量很好的解决了这个问题，从概念上讲，词向量是一种将维数为词表大小的高维空间中的向量投影到一个维数低得多的连续向量空间后得到的表示，由于解决了 One-hot Representation 存在的问题，词向量正作为一种优秀的语言表示方式在自然语言处理领域发挥着重要作用。

最近几年, 随着机器学习和深度学习在自然语言处理领域的不断兴起和广泛应用, 融入语言学知识源信息的词向量表示和基于神经网络的词向量表示开始应用在许多自然语言处理的任务中。其中单语词向量 (Monolingual word embedding) 因其优异的性能被广泛应用在各个领域中，比如情感分析 (Socher et al. 2013)[1], 依存分析 (Dyer et al. 2015[2], Guo et al. 2015[3]) 等等。

然而，这些单语模型的优异性能通常被限制在训练模型所用的语言中。而资源的可用性，训练数据的规模和性能评价基准等诸多因素导致了大多数模型都是基于英语的，而在其他许多资源比较少的语言上由于资源的稀缺则训练不出优异的模型。为了解决这个问题并对语言学习环境进行调整，人们希望利用现有的多语知识来为当前的单语模型提供处理其他语言的能力。

诚然，机器翻译可以解决这个问题，我们可以将一个语言翻译到另一个语言，然后再进行需要的任务。然而我们完全不必采用这种复杂的做法，我们可以将两个语言的词都投到一个共有的子空间中，使语义相近的词拥有相近的表示，就像图 1-1中展示的那样。我们的目标是在所有语言的单词之间学习一个共享的向量空间。如果拥有了这样的矢量空间，我们可以用任何语言对数据进行训练。通过将一种语言的实例投射到这个空间中，我们的模型同时获得了以所有其他语言执行预测的能力，这就是跨语言词向量的由来。

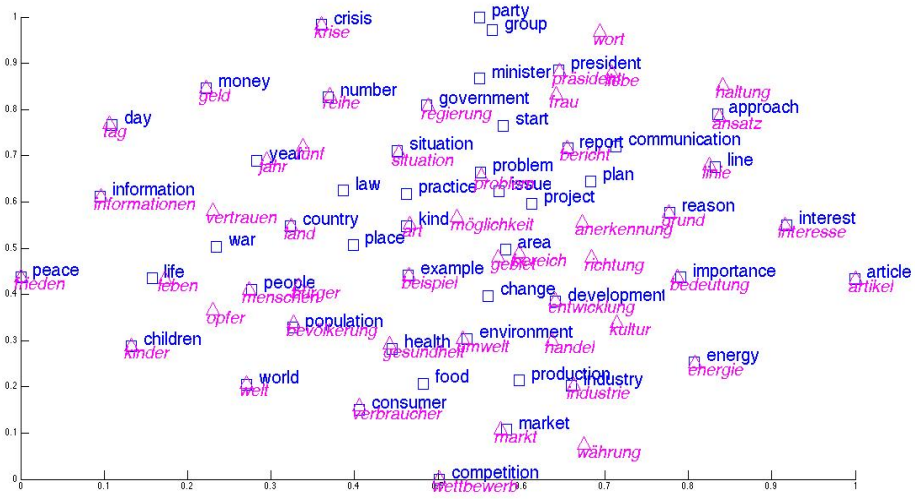

图 1-1　两个语言的共享词向量空间 (Luong et al., 2015[4])

1.2 国内外研究现状

跨语言词向量 (cross-lingual word embeddings) 作为对单语词向量的一种扩展，它不仅提高了词向量在单语任务上的性能 (Faruqui and Dyer 2014[5])，而且也为很多跨语言任务提供了新鲜血液，比如机器翻译 (machine translation)(Uszkoreit et al. 2010[6], Mikolov et al. 2013b[7], Zhang et al. 2014[8]), 跨语言文本分类 (cross-lingual document classification)(Wan 2009[9], Klementiev et al. 2012[10], Shi et al. 2015[11],Upadhyay et al. 2016[12]), 跨语言依存分析 (cross-lingual dependency parsing) (McDonald et al. 2013[13], Gouws et al. 2015[14])。

近几年来，人们提出了很多模型来生成跨语言的词向量，以求在各种任务上获得更好的性能。从最优化的角度来讲，这些模型大致可以分为以下两类：

- 对于第一类模型 (Mikolov et al. 2013b[7], Faruqui and Dyer 2014[5]), 先根据每个单语预料训练出一个单语模型（词向量），这一步是离线运行的，无需任何双语信息；然后再单独一步将前面获得的单语词向量融合多语信息进而转换成多语词向量；

- 对于第二类模型 (Hermann and Blunsom 2014[15], Shi et al. 2015[11], Gouws et al. 2015[14], Luong et al. 2015[4])，来自多个语言的单语语料被用来联合训练 (jointly training)，使用跨语言的知识作为约束条件，直接训练得到多语词向量。相比于第一类，第二类模型通常是低计算复杂度、容易扩展到大的数据集上，且有着相对

不错的性能。

除了模型外，训练需要的平行语料也各有不同，从平行语料的代价由大到小可以分为:

• 词对齐语料 (Word-aligned data)，词与词之间是对其的，通常被用来训练机器翻译模型，是所有平行语料中获得代价最高的语料。

• 句对齐语料 (Sentence-aligned data)，句子间是对齐的，大多数数模型使用的是欧洲议会语料，代价仅次于词对齐语料

• 文件对其语料 (Document-aligned data)，文件与文件间是对齐的，文件可以根据话题对齐 (如 Wikipedia)，也可以根据类别/标签对齐 (比如情感分析和多分类数据集)

• 词典 (Lexicon)，由双语或者多语言互译词对组成的集合，比较廉价，容易扩展。

• 无平行语料 (No parallel data)，只利用单语语料进行学习，跨语言实现 zero-shot learning。

本文重点介绍和改进的模型 BiCCA(Faruqui and Dyer 2014[5]) 属于上述模型分类中的第一类，使用廉价的双语词典进行训练，总体来说是一个比较 ‘‘节能’’ 的模型。BiCCA 的主要思想是首先假设不同语言两个可以互译的词之间是有着比较强的线性关系的 (Mikolov et al. 2013b[7])，此模型使用典型相关分析 (CCA) 来学习这种线性关系，并将这种关系融合到原来的单语词向量中生成多语词向量。然而在我们的实践中，我们发现事实上不同语言互译词之间是很多非线性关系的，这些非线性关系显然是不能通过 CCA 这个线性模型来捕获，因此我们提出使用更加强力的核典型相关分析 (KCCA,Akaho 2006[16]) 来同时捕获语言间的线性和非线性关系，进而得到一个新的跨语言语义表示方式。

1.3 文章主体结构

本文的主体结构主要安排如下：

第一章：绪论，主要介绍课题的背景、国内外在该方向的研究现状，介绍并指出了现有模型存在的问题，最后给出了本文的结构。

第二章：现有模型，主要介绍现有模型 BiCCA 的原理，生成词向量的过程，以及存在的问题。

第三章：新模型介绍，主要介绍我们提出的新模型的数学原理，模型实现以

及训练过程。

第四章：评估实验，主要介绍我们用到的三个评估实验 — 两个上游评估实验，一个下游评估实验。

第五章：实验，主要介绍实验的语料、模型的参数、实验的其他细节以及评估实验的结果。

第六章：分析，主要对评估实验的结果、错误以及模型做了系统的分析。

第七章：相关工作，主要介绍近年来相关技术的发展过程。

最后一部分是总结与未来工作，对完成的模型和实验结果进行总结，并对未来的研究工作指出方向。

第 2 章 现有模型

2.1 现有双语词向量生成模型 (BiCCA)

在本章中，我们将简要介绍 BiCCA 模型 (Faruqui and Dyer 2014[5])，它利用 CCA 来学习语言间的线性关系并将离线训练得到的单语词向量转换成多语词向量。通过 BiCCA，我们为两个源语言生成了新的带有双语知识的词向量，而这些词向量通常在很多任务中相比于单语词向量有着更好的性能。

假设我们已经具有两个源语言的单语词向量，表示为 $\Sigma \in \mathbb{R}^{n_1 \times d_1}$, $\Omega \in \mathbb{R}^{n_2 \times d_2}$，以及两个语言上的一个双语词典，BiCCA 通过构造好的双语词向量词典训练出一对线性变换，并通过这个线性变换将原来的单语词向量矩阵 Σ, Ω 投影到一个新的共享的向量空间。我们希望生成的词向量在共享的向量空间中拥有如下两点性质：（1）同一语言中语义相近的词在空间中也是相近的。（2）不同语言间语义相近的词在空间中也是相近的。

2.1.1 BiCCA 数学原理

假设 $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ and $\Omega \in \mathbb{R}^{n_2 \times d_2}$ 分别对应两个语言的单语词向量矩阵，其中 n_1, n_2 代表词典的大小，d_1, d_2 代表词向量的维度。我们首先构造训练词向量矩阵 Σ', Ω'，构造方法是选择原词向量集合 Σ, Ω 的子集 $\Sigma' \subset \Sigma$，$\Omega' \subset \Omega$, 满足 $|\Sigma'| = |\Omega'|$；且对任意 i，满足 (Σ'_i, Ω'_i) 是一对可以互译的词。使用得到的训练数据训练 CCA 模型，可以得到两个线性变换（投影）矩阵 P_Σ, P_Ω：

$$P_\Sigma, P_\Omega = CCA(\Sigma', \Omega') \tag{2-1}$$

其中 $P_\Sigma \in \mathbb{R}^{d_1 \times k}$，$P_\Omega \in \mathbb{R}^{d_1 \times k}$, $k \leq min\{d_1, d_2\}$, CCA 将在本章的下一部分做出介绍。利用训练得到的投影向量矩阵，可以将源词向量矩阵投影到新的共享的向量空间中：

$$\Sigma^* = \Sigma P_\Sigma, \Omega^* = \Omega P_\Omega \tag{2-2}$$

其中 $\Sigma^* \in \mathbb{R}^{n_1 \times k}, \Omega^* \in \mathbb{R}^{n_2 \times k}$ 是用双语知识“丰富”后得到的双语词向量。整个使用 BiCCA 生成双语词向量的过程可以用图 2-1表示。

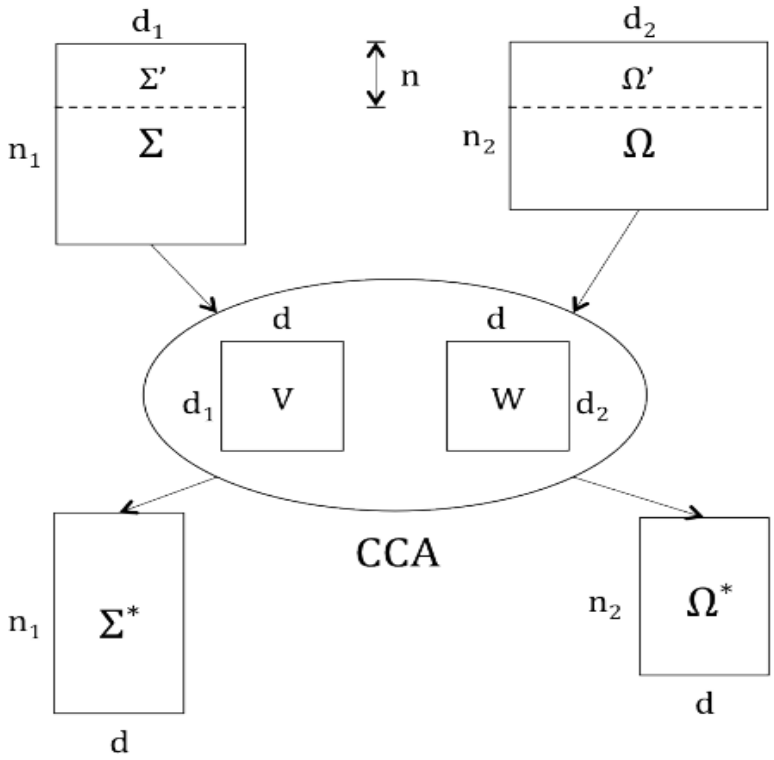

图 2-1　BiCCA 先构造 Σ', Ω' 作为输入给 CCA，得到一对线性变换 $v \in \mathbb{R}^{n_x}, w \in \mathbb{R}^{n_y}$，再利用线性变换 v, w 将原来的单语词向量 Σ, Ω 转换为双语词向量 Σ^*, Ω^*。

2.2 典型相关分析 (CCA)

一种典型的用来学习跨语言表示的方法就是典性相关分析 (CCA,Hotelling 1936[17])，CCA 是一个利用投影变量对之间的相关关系来反映两个多维随机变量之间的整体相关性的多元统计分析方法。CCA 寻求为每个多维随机变量寻求一个投影变量，使得两个多维随机变量投影后的相关系数最大，这些新投影的向量的维度 k 小于等于原来的两个多维随机变量的维度。

2.2.1 CCA 原理简介

图 2-2 简要的介绍了 CCA 的主要工作原理①。对于多维随机变量 $X = (X_1 X_2 .... X_m), Y = (Y_1 Y_2 .... Y_n)$，CCA 寻找两个投影向量 a, b, 分别与 X, Y 相乘得到

$$u = a^T X, v = b^T Y \tag{2-3}$$

然后使用 u 与 v 的 Pearson 相关系数

① 更多关于 CCA 的内容，请参照 Relations Between Two Sets of Variates (Hotelling 1936[17])

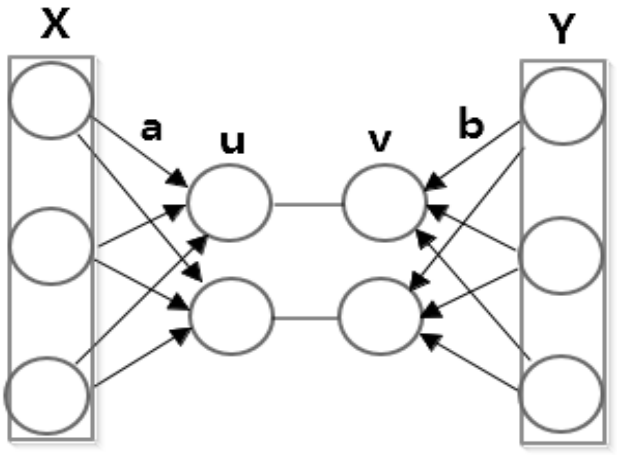

图 2-2　CCA 寻求一对线性变换 $a \in \mathbb{R}^{n_x}$, $b \in \mathbb{R}^{n_y}$ 以使得 $u = a^T X, v = B^T Y$ 间的相关系数 $\rho = corr(u, v)$ 最大化。

$$\rho(u,v) = \frac{cov(u,v)}{\sqrt{Var(u)}\sqrt{Var(v)}} = \frac{E[(u-\mu_x)(v-\mu_y)]}{\sqrt{Var(u)}\sqrt{Var(v)}} \tag{2-4}$$

来度量 X 和 Y 的关系, 我们期望寻求一组最优的解 a 和 b 使得相关系数 $\rho(u, v)$ 最大, 这里的 μ_x, μ_y 分别是 u 和 v 的均值。

随机变量 $u = a^T X$和$v = b^T Y$ 是第一对典型变量。然后寻求一个依然最大化相关但与第一对典型变量不相关的向量；这样就得到了第二对典型变量, 这个步骤会进行 $min\{m, n\}$ 次，m, n 分别是 X, Y 的维度。

因此 CCA 的主要工作就是求 a 和 b 两个投影向量，求得了 a,b，也就求得了典型变量。

2.2.2 CCA 数学求解过程

给定两组向量 X 和 Y，X 维度为 d_1, Y 维度为 d_2。构造矩阵 T 并形式化表示如下:

$$T = \begin{bmatrix} X \\ Y \end{bmatrix}, E(T) = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{2-5}$$

其中 Σ 是 T 的协方差矩阵，Σ_{11} 是 X 自己的协方差矩阵，也就是 X 的方差矩阵,Σ_{22} 是 Y 的方差矩阵，Σ_{12} 是 X 与 Y 的协方差矩阵，Σ_{21} 是 Σ_{12} 的转置矩阵。

由公式 2-3、公式 2-5可以求得:

$$Var(u) = Var(a^T X) = \frac{1}{N}\sum (a^T X_i - a^T \mu_x)^2 = a^T \frac{1}{N}\sum (X_i - \mu_x)^2 a = a^T \Sigma_{11} a \tag{2-6}$$

同理还可以求得:

$$Var(v) = b^T\Sigma_{22}b, Cov(u,v) = a^T\Sigma_{12}b \tag{2-7}$$

因此 u, v 间的相关系数可以表示为:

$$\rho(u,v) = \frac{cov(u,v)}{\sqrt{Var(u)}\sqrt{Var(v)}} = \frac{a^T\Sigma_{12}b}{\sqrt{a^T\Sigma_{11}a}\sqrt{b^T\Sigma_{22}b}} \tag{2-8}$$

容易观察到上面的公式 2-8中 $a = ka, b = mb$ 时结果不变, 因此原问题等价于加上约束:

$$Var(u) = Var(v) = 1 \tag{2-9}$$

因此求解 CCA 的参数等价于解决以下最优化问题:

$$Maximize : a^T\Sigma_{12}bSubjectto : a^T\Sigma_{11}a = 1, b^T\Sigma_{22}b = 1 \tag{2-10}$$

构造 Lagrangian 等式:

$$L = a^T\Sigma_{12}b - \frac{\lambda}{2}(a^T\Sigma_{11}a - 1) - \frac{\theta}{2}(b^T\Sigma_{22}b - 1) \tag{2-11}$$

求偏导，令导数得 0:

$$\Sigma_{12}b - \lambda\Sigma_{11}a = 0 \tag{2-12}$$

$$\Sigma_{21}a - \theta\Sigma_{22}b = 0 \tag{2-13}$$

令公式 2-12左乘 a^T, 公式 2-13左乘 b^T, 再由 $a^T\Sigma_{11}a = 1, b^T\Sigma_{22}b = 1$ 得:

$$\lambda = \theta = a^T\Sigma_{12}b \tag{2-14}$$

也就是说 λ 就是 $corr(u,v)$，只要求出最大的 λ 即可。

由公式 2-13导出 b，再带入公式 2-12中，结合公式 2-14$\lambda = \theta$ 可得:

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}a = {}^2a \tag{2-15}$$

因此只需求出矩阵 $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ 的最大的特征值和对应的特征向量就能求得 λ 和 a，同理 b 也可以求得。

至此我们终于完成了第一对典型变量对应的投影变量的求解，如果需要求解更多的投影变量，只要求解该矩阵的其他的特征值和特征向量即可。

第 3 章 我们提出的模型

尽管我们在第二章所介绍的（BiCCA）已经在跨语言词向量生成任务中取得了不错的效果，但我们发现（后面会给出例子），语言之间不是只有线性关系，还存在这一定量的非线性关系，而这是 BiCCA 所捕捉不到的。因此我们提出使用核典性相关分析 (KCCA) 来捕获语言间的非线性关系，获得新的跨语言表示。

3.1 核典型相关分析 (KCCA)

在这一节中，我们主要介绍 KCCA 模型。KCCA 模型主要是用来解决 CCA 不能捕捉非线性关系的问题。

我们先来解释 KCCA 为什么可以捕捉非线性关系，图 3-1生动地解释了了 KCCA 为什么可以捕捉非线性关系。在图 3-1中，左边展示的是随机变量 X, Y 的原联合分布情况，显然 X, Y 间并不是简单的线性关系。中间部分展示的是一对非线性映射函数 $\hat{f}(x), \hat{g}(y)$，这里我们使用的是 Gaussian RBF kernel $k(x, y) = exp(-\frac{1}{2\sigma^2}(x-y)^2)$。右边展示的是通过映射函数映射后的结果。不难看出，本来是非线性关系的 X, Y 经过映射函数转换以后，形成的 $\hat{f}(X), \hat{g}(Y)$ 间的关系是线性的，

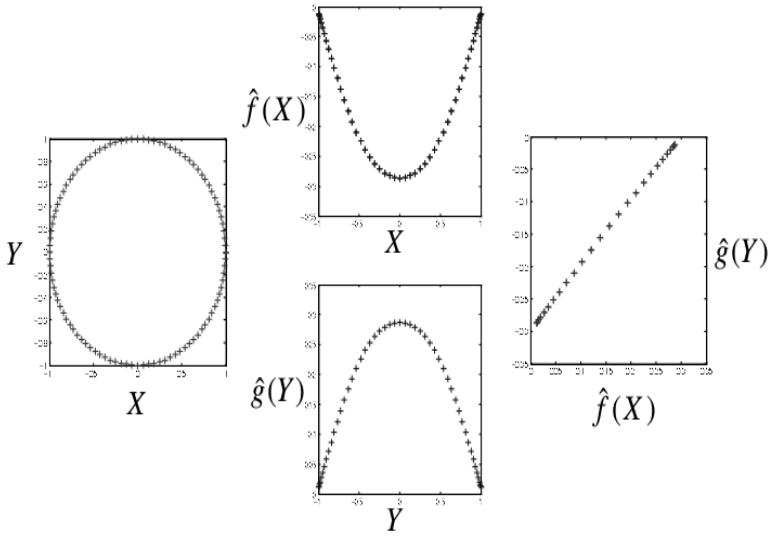

图 3-1　KCCA 通过两个非线性的映射函数将原来的非线性关系转换为线性关系，再通过 CCA 去捕获转换后的线性关系。左: X, Y 的原联合分布情况. 右: 经过转换后得到的 $\hat{f}(X), \hat{g}(Y)$ 的联合分布. 中间: 非线性映射函数 $\hat{f}(x), \hat{g}(y)$。

而这可以用前面介绍的 CCA 来捕捉。由此可以说明 KCCA 是可以捕获非线性关系的。

接下来的部分，我们将介绍 KCCA 的数学原理及求解过程[①]。核函数版本的 CCA 首先利用核函数 Φ_x, Φ_y 将初始数据 X, Y 映射到 Hilbert 空间中，核函数可以表示为:

$$\Phi : X = (X_1 X_2 \cdots X_n) \rightarrow \Phi(X) = (\phi_1(X), \phi_2(X) \cdots \phi_N(X)), \text{n<N.}$$

通常核函数都是非线性的映射函数。一般来时经过核函数转换后的关系都会由原来的非线性转化为线性关系，此时就可以用 CCA 进行捕捉。

假设多维随机变量 X, Y，我们要捕获 X, Y 间的非线性关系，KCCA 首先将源数据 X, Y 映射到高维 Hilbert 空间中，得到 $\Phi_x(X) \in H_x, \Phi_y(Y) \in H_y$。我们的目标是寻找一对投影向量 $a \in H_x, b \in H_y$，使得投影后得到的 u, v:

$$u = a^T \Phi_x(X) \tag{3-1}$$

$$v = b^T \Phi_y(Y) \tag{3-2}$$

之间的相关系数 $\rho = corr(u, v)$ 最大，整个过程可以用图 3-2表示。

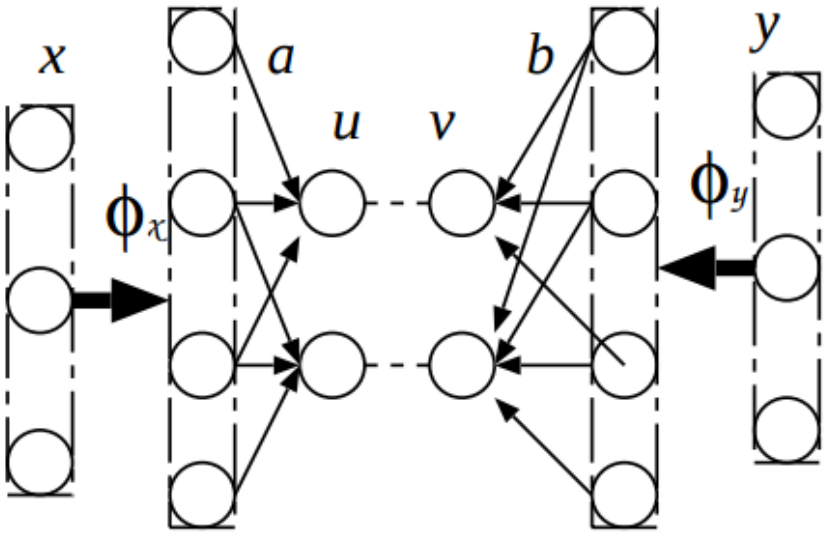

图 3-2　KCCA 寻找高维 Hilbert 空间中的两个投影向量 $a \in H_x, b \in H_y$，使得 $u = a^T \Phi_x(X), v = b^T \Phi_y(Y)$ 间的相关系数 $\rho = corr(u, v)$ 最大。

① 关于更多细节，请参照 Statistical Consistency of Kernel Canonical Correlation (Fukumizu et al. 2007[18])

3.1.1 KCCA 求解过程

类似于 CCA，我们也希望直接求解出一对投影向量 $a \in H_x, b \in H_y$，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 a, b 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 X^TX 型，进而替换成核矩阵 K，而使用上面的做法不能出现 X^TX 型，因此很难利用核矩阵，求解比较复杂。

因此我们考虑以对偶形式解决这个问题。重新定义 $P = \begin{bmatrix} \Phi(X_1)^T \\ \Phi(X_2)^T \\ \vdots \\ \Phi(X_m)^T \end{bmatrix}$，则 $P^T = \begin{bmatrix} \Phi(X_1)\ \Phi(X_2) \cdots \Phi(X_m) \end{bmatrix}$，同理构造 $Q = \begin{bmatrix} \Phi(Y_1)^T \\ \Phi(Y_2)^T \\ \vdots \\ \Phi(Y_m)^T \end{bmatrix}$(这里 X,Y 的样本数量是相同的)

原投影向量 a, b 可以表示成训练样本的一个线性组合[①]：

$$a = P^T\alpha. \tag{3-3}$$

$$b = Q^T\beta. \tag{3-4}$$

此时，u, v 可以表示成:

$$u = \alpha^T P, v = \beta^T Q \tag{3-5}$$

此时，我们需要最大化的相关系数：

$$\rho(u, v) = \frac{cov(u, v)}{\sqrt{Var(u)}\sqrt{Var(v)}} \tag{3-6}$$

其中：

$$Var(u) = a^T var(P)a = \alpha^T Pvar(P)P^T\alpha \tag{3-7}$$

$$Var(v) = b^T var(Q)b = \beta^T Qvar(Q)Q^T\beta \tag{3-8}$$

① 这个已经由 (Fukumizu et al. 2007[18]) 证明过, 通过将 a 和 b 表示成这个形式, 可以使用核-trick。

$$Cov(u,v) = a^T cov(P,Q)b = \alpha^T Pcov(P,Q)Q^T\beta \tag{3-9}$$

进行下一步前，先对 P,Q 进行正则化处理, 即 $\sum xi = 0$，这样就满足 $cov(P,Q) = \frac{1}{N-1}P^TQ$。给定核函数 Φ_x and Φ_y, 我们令 K_p 和 K_q 表示对应的核矩阵, 可以表示为 $K_p = PP^T, K_q = QQ^T$。带入上面的三个公式，可以得到:

$$Var(u) = \alpha^T Pvar(P)P^T\alpha = \alpha^T PP^T PP^T\alpha = \alpha^T K_p^2\alpha \tag{3-10}$$

$$Var(v) = \beta^T Qvar(Q)Q^T\beta = \beta^T QQ^T QQ^T\beta = \beta^T K_q^2\beta \tag{3-11}$$

$$Cov(u,v) = \alpha^T Pcov(P,Q)Q^T\beta = \alpha^T PP^T QQ^T\beta = \alpha^T K_pK_q\beta \tag{3-12}$$

因此:

$$\rho(u,v) = \frac{cov(u,v)}{\sqrt{Var(u)}\sqrt{Var(v)}} = \frac{\alpha^T K_pK_q\beta}{\sqrt{\alpha^T K_p^2\alpha}\sqrt{\beta^T K_q^2\beta}} \tag{3-13}$$

容易观察到上面的公式 3-13中 $\alpha = k\alpha, \beta = m\beta$ 时结果不变，因此原问题等价于加上约束:

$$Var(u) = Var(v) = 1 \tag{3-14}$$

因此求解 KCCA 的参数等价于解决以下最优化问题:

Maximize: $\alpha^T K_pK_q\beta$

Subject to: $\alpha^T K_p^2\alpha = 1, \beta^T K_q^2\beta = 1$

构造对应的 Lagrangian 等式:

$$L = \alpha^T K_pK_q\beta - \frac{\lambda}{2}(\alpha^T K_p^2\alpha - 1) - \frac{\theta}{2}(\beta^T K_q^2\beta - 1) \tag{3-15}$$

类似于 CCA 的求解过程，最同样可以得到一个求矩阵特征值的方程:

$$K_pK_p\alpha - \lambda^2 K_pK_p\alpha = 0 \tag{3-16}$$

即:

$$I\alpha = \lambda^2\alpha \tag{3-17}$$

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

3.1.2 正则化的 KCCA

上一节中介绍了基本的 KCCA 求解过程，指出 KCCA 是拥有解析解的。然而上一节中介绍的 KCCA 是过拟合的模型，我们需要继续添加正则项来避免过拟合

问题。

一般来说，有两种常见的正则化的方法，一种是对 Lagrange 等式加入正则项，一种是对限制条件做一下修改，我们这里将主要介绍第一种做法，这也是我们模型实现中所采用的算法。

第一种方法是在 Lagrange 等式后加入模型的复杂度，具体写为 $\frac{\eta}{2}(\| \alpha \|^2 + \| \beta \|^2)$，因此对应的 Lagrange 等式变为：

$$L = \alpha^T K_p K_q \beta - \frac{\lambda}{2}(\alpha^T K_p^2 \alpha - 1) - \frac{\theta}{2}(\beta^T K_q^2 \beta - 1) + \frac{\eta}{2}(\| \alpha \|^2 + \| \beta \|^2) \tag{3-18}$$

同样令 Lagrange 函数的偏导数为 0，得：

$$K_p K_q \beta = (\lambda K_p^2 + \eta I)\alpha \tag{3-19}$$

$$K_q K_p \alpha = (\theta K_q^2 + \eta I)\beta \tag{3-20}$$

省略中间计算步骤 (详见文章附录)，最后化简可以求得：

$$(K_p + \eta I)^{-1} K_q (K_q + \eta I)^{-1} K_p \alpha = \lambda^2 \alpha \tag{3-21}$$

此时，不难发现 α, β 就对应着左边矩阵的特征向量，只要求解出矩阵的特征向量，问题就能够解决[①]。

第二种方法是对限制条件作修改，修改后的模型对应的最优化问题是：

Maximize: $\alpha^T K_p K_q \beta$

Subject to: $\begin{aligned}(1-\sigma)\alpha^T K_p^2 \alpha + \sigma \alpha^T K_p \alpha = 1 \\ (1-\sigma)\beta^T K_q^2 \beta + \sigma \beta^T K_q \beta = 1\end{aligned}$

这种方法求解后也会得到类似于方法一的结果，我们这里就不再花笔墨去详细介绍了。

至此，我们完成了第一对典型变量对应的投影变量的求解，如果需要求解更多的投影变量，类似于 CCA，只要求解该矩阵的其他的特征值和特征向量即可。

3.2 使用 KCCA 生成跨语言的词向量

本节中我们将会描述如何利用 KCCA 去学习一个跨语言的语义表示。我们的初始条件与前人的工作相同，也是两个语言的词向量矩阵，表示为 Σ, Ω，还有一个互译的词典。首先，我们使用 KCCA 去学习两个语言词向量矩阵间的关系，通过训练得出一对非线性的转换矩阵；然后我们再利用这个非线性的转换矩阵将原来

① η 是已知的变量，I 是单位矩阵

的单语词向量矩阵投影到新的共享的向量空间，也就得到了跨语言的词向量。这个过程在图 3-3中做了展示。

令 $\Sigma \in \mathbb{R}^{n_1 \times d_1}$ 和 $\Omega \in \mathbb{R}^{n_2 \times d_2}$ 代表两个语言的词向量，其中 n_1, n_2 代表词向量的个数，d_1, d_2 代表词向量的维度，训练矩阵 Σ', Ω' 按照下面的方法构造：$\Sigma' \subset \Sigma$，$\Omega' \subset \Omega$，且两者词向量的个数相同，对于任意第 i 对词，Ω_i'，Σ_i' 是互译的。

我们使用 KCCA 去最大化训练矩阵 Σ' ，Ω' 间的相关系数 ρ 并输出两个参数 α, β，也就是所求的投影变量。

$$\alpha, \beta = KCCA(\Sigma', \Omega') = \arg\max_{\alpha,\beta} \rho(a^T \Phi_x(\Sigma'), b^T \Phi_y(\Omega')) \tag{3-22}$$

其中公式 3-22中的 a, b 可以像公式 3-3，公式 3-4一样表示成含有 α, β 的等式。使用得到的投影变量 α, β，我们可以像公式3-5一样将整个源语言的词向量投影到新的向量空间中，并作为新的跨语言词向量。这个过程可以用矩阵的形式简化表示如下：

$$\alpha_k, \beta_k = KCCA(\Sigma', \Omega') \tag{3-23}$$

$$\Sigma_k^* = P_\Sigma P_{\Sigma'}^T \alpha_k, \Omega_k^* = Q_\Omega Q_{\Omega'}^T \beta_k \tag{3-24}$$

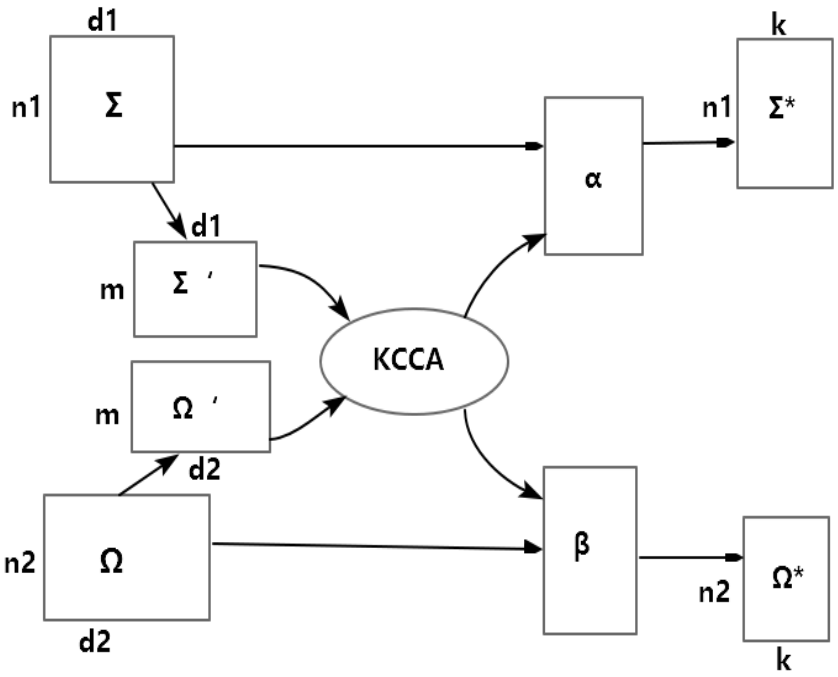

图 3-3　使用 KCCA 生成双语词向量

其中，P_Σ 矩阵的每一行是向量 $\Phi_x(X_i)$[①], $X_i \in \Sigma$；Q_Ω 矩阵的每一行是向量 $\Phi_y(Y_i), Y_i \in \Omega$；$P_{\Sigma'}, Q_{\Omega'}$ 类似于 P_Σ, Q_Ω. 我们的实验中选取了 top-k 相关的维度，α_k, β_k 就是对应的 k 对投影向量。$\Sigma_k^* \in \mathbb{R}^{n_1 \times k}, \Omega_k^* \in \mathbb{R}^{n_2 \times k}$ 表示两个语言新的具有双语知识的词向量表示。

至此，我们解释了如何使用 KCCA 去生成跨语言的词向量。

① Φ_x, Φ_y 是 §3.1 中定义的核函数

第 4 章 词向量的评估

Upadhyay et al. (2016)[12] 对四种当前比较流行的词向量生成模型生成的词向量进行了比较系统的评估实验，我们仿照他的实验方法来对我们的模型生成的词向量进行评估实验。我们主要选取了以下三个任务来评估我们的词向量。

- 英语的单语词相似度实验（Monolingual word similarity for English）
- 跨语言词典生成实验（Cross-lingual dictionary induction）
- 跨语言文本分类实验（Cross-lingual document classification）

前两个任务主要用来评估跨语言训练究竟能给词向量带来多大的提升，最后一个任务用来测试基于跨语言词向量的模型在不同语言间的迁移能力①。

4.1 单语评估实验

我们首先评估我们的模型是否能够提升英语词向量的质量。单语评估任务可以衡量训练得到的词向量在语义上是否符合人们的直觉。

4.1.1 单语词相似度评估

词相似度数据集（Word Similarity dataset）中主要包含人们对每一对词的相似度的等级评分。词相似度评估实验可以衡量机器生成的词向量和人工生成的词相似度数据集中的词在语义上的相似程度。评估的结果是先计算两个词词向量之间的余弦相似度 (cosine similarity)，然后再根据 Spearman 相关系数 (Myers and Well, 1995[19]) 计算人们的评分和前面得到的余弦相似度的相关系数，最后给出整体的结果。

我们使用 SimLex dataset for English (Hill et al. 2014[20]) 数据集，它包含 666 个名词对，222 个动词对，111 个形容词对。SimLex 不同于另一个常用的数据集 WordSim-353 (Finkelstein et al. 2001[21])，Simlex 只关注词的相似度，而不关心词与上下文间的联系，这对我们实验来说是减少了干扰因素。

4.1.2 QVEC 评估

我们使用 QVEC-CCA（Waleed Ammar et 2016[22]）对词向量进行评估。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-

① 为了保证公平科学对比，两个模型均以 200 维的训练数据训练, 产生 100 维的跨语言词向量。

CCA 将 QVEC 和 CCA 相结合，先通过最大化训练词向量矩阵 X 和语言本身的‘‘特性’’矩阵 S 间的线性关系训练模型，然后通过计算两者的线性相关系数来作为衡量词向量质量的指标。QVEC 产生的分数可以定量衡量一个词的词向量表达原词语言特性的能力。对于 QVEC-CCA 评估任务，我们同样在 SimLex-999 数据集上进行。

为了确认我们模型相对于原模型的提升程度，我们两种评估中均提供了 P-value(Steiger, 1980[23]) 测试，对于本评估实验，如果 P-value 值 $P < 0.15$，则我们宣称我们的模型相对于原模型有重大提升。

4.2 跨语言词典生成

在这个任务中，我们主要考察非线性模型是否可以在跨语言词典生成任务中相比于原模型获得性能提升。跨语言词典生成任务 (Vulić and Moens 2013[24]; Gouws et al. 2015[14]; Mikolov et al. 2013b[7]; Upadhyay et al. 2016[12]; Zhang et al. 2016[25]) 可以评估模型生成的跨语言词向量在探寻不同语言间语义相近的词任务上的能力。

我们参照 Upadhyay 的做法, 从 Open Multilingual WordNet data 数据集 (Bond and Foster,2013[26]) 中抽取出 gold 词典进行评估实验，该数据集中包含 26 种语言的同义词对齐集合，并有着 90% 的准确度。具体地，我们首先筛掉那些出现频率低于 1000 的词，之后，对于每一组同义词集合 $s1 = \{k_1, k_2, \cdots\}, s2 = \{g_1, g_2, \cdots\}$，我们将所有不重复的组合 (k_i, g_j) 添加到我们的 gold 词典中，使用这样的方法，我们为 en-fr, en-de, en-ar, en-ru 和 en-zh 五个语言对各自生成了单词个数为 1.5k, 1.4k, 1.5k, 1.4k 和 1.6k 的词典。

本实验中，我们统计 $Top-k(k \in \{1, 10, 50\})$ 准确度，对于 gold 词典的每一对词 (e, f)，我们先在向量空间中找到 e，然后找到其他语言中距离 e 最近的 k 个词，构成一个列表 L，我们判断 f 是否在列表 L 中，综合所有词对，我们得到了测试集的正确率。

4.3 跨语言文本分类

跨语言文本分类（Cross-lingual document classification）任务可以评估不同语言的跨语言词向量能否在语义上保持一致性。

我们参照 Upadhyay et al. (2016)[12] 在跨语言文本分类任务中的做法进行我们的评估实验，我们将评估的语言对扩展为 en-fr, en-de, en-ar, en-ru 和 en-zh 五个语

言对。此外，对于评估语料，我们选用的不是传统的 RCV2 (Lewis et al. 2004[27]) 语料，而是另一个比较常用的 WIT TED (Cettolo et al. 2012[28]) 语料。TED 语料的来源主要是不同语言的人们对每个 TED 的评价，经过处理后形成了 15 个话题的训练和测试语料，相比于 RCV2 的四个话题，TED 对模型的考察更为全面。

具体地，对于每一个语言对 (l_1, l_2)，我们使用由语言 l_1 的词向量生成的文章向量去训练一个分类模型，之后在 l_2 的文章向量上进行测试（反之亦然）。根据前人的工作，我们的文章向量由词向量加权平均计算而来，权重用的是 TF-IDF 值①。同样延续前人的工作，我们采用一个多分类的平均感知机 (averaged perceptron,Freund and Schapire, 1999[30]) 迭代 10 轮作为文本分类器。

我们引用了 Hermann and Blunsom[15] 在 TED 语料上训练的 MT 系统的结果作为我们实验的 baseline。他们的系统中先使用 CDEC（Dyer et al. 2010[31]）系统进行文档的翻译，之后训练一个朴素贝叶斯模型对文档进行分类。

通过仅在一个语言上进行训练，而在另一个语言上不做其他监督学习直接进行测试，跨语言文本分类任务可以衡量跨语言的词向量促进模型迁移任务中的能力。

① TF-IDF (Salton and Buckley 1988[29]) 由 TED 语料中每个语言的所有文档计算而来。

第 5 章 实验

在本章中我们将运行第四章中描述的任务来评估我们的词向量并给出基本的实验结果。

5.1 实验数据

为了实验更具代表性，我们从不同语系中选取了 5 种语言作为代表进行实验，分别选择了英语-德语 (en-de)、英语-法语 (en-fr)、英语-俄语 (en-ru)、英语-阿拉伯语 (en-ar)、英语-汉语 (en-zh)。对于英语、法语、德语，单语语料主要从 Europarl[①]中获得；对于阿拉伯语、俄语、汉语，单语语料主要来自 Leipzig[②]语料库。

为了构造训练词典，我们使用了 Europarl 的平行语料。首先，对于词对 $(a, b), a \in l_1, b \in l_2$，如果 a, b 满足在平行语料中 a 被翻译成 b 的次数最多且 b 被翻译成 a 的次数最多，则 (a, b) 符合要求。然后，我们选取最高频的词对加入我们的训练词典中。对于 Europarl 不存在对应平行语料的语言 (阿拉伯语，俄语，汉语)，我们使用谷歌翻译系统翻译了英语语料中最常见的 20 万个词，将翻译结果与原词形成词对构成训练词典。在实际训练过程中，为了平衡训练速度与性能，我们使用大小为约 7000 对词的词典进行训练。

5.2 实验参数

为了方便对比，我们采用与 Upadhyay 相同的实验设置进行实验。首先，对于单语词向量，我们使用 Word2vec[③]的 skip-gram 模型使用 negative sampling，窗口大小设置为 5 在单语语料上训练得来; 对于训练词典，我们使用 §5.1 中描述的方法获得。我们使用 $k = 0.5$ 作为维度系数，这样我们以 200 维的词向量训练，产生 100 维的词向量进行评估。另外，所有词向量在进行下一步的应用前都会进行单位化。

CCA 模型本身比较简单，不涉及模型的参数。下面主要介绍 KCCA 模型的一些参数：对于核函数，我们为两种语言都选择了 RBF 核[④]$k_1(x_i, y_i) = exp(-\gamma(\| x_i - y_i \|^2)$，$k_2$ 与 k_1 类似，核函数参数的调节范围是 $[10^{-1}, 10^{-5}]$。对于公式3-21中

① http://www.statmt.org/europarl/

② http://wortschatz.uni-leipzig.de/en

③ code.google.com/p/word2vec

④ 我们同样尝试了多项式核 (poly-kernel)，性能上并没有提升。

的正则化因子，参数的调节范围是 $[10, 10^{-6}]$。

对于每个评估实验，我们进行 5 等分的交叉验证实验，然后选取综合性能最好的超参数作为模型参数。

5.3 实验结果

5.3.1 单语词相似度

表 5-1中展示了我们在单语词相似度评估任务 (§4.1.1) 上的主要结果。我们比较了单语词向量 (Mono)、基于 CCA 的模型的跨语言词向量 (BiCCA)、基于 KCCA 的模型的跨语言词向量 (BiKCCA) 三种模型的结果，我们选用单语词向量的表现作为 base-line。如果两个模型的结果按照 Steiger 的方法计算得到的 P-value 值 (p<0.15)，我们则宣称取得了有效提升。

通过观察表 5-1中的结果，我们不难发现：

- 两个双语模型的结果都比单语模型的结果要好。
- BiKCCA 相对于 BiCCA 几乎在所有语言对上都取得了性能的提升。

对于第一点，这符合我们的预期，也验证了 Faruqui 的结论：融合了跨语言知识的双语词向量相比于单语词向量在语义上有着更高的质量。

对于第二点，这样的结果意味着语言之间的非线性关系可以进一步提升词向量在语义上的质量。

除此以外我们还发现，对于本质上差异比较大的语言，如英语-俄语、英语-阿拉伯语、英语-汉语，BiKCCA 相比于 BiCCA 有着更显著的提升。

表 5-2中展示了我们在 QVEC-CCA 评估任务 (§4.1.2) 上的主要结果。我们同

表 5-1　词向量的单语词相似度评估结果

L_1	L_2	Mono	BiCCA	BiKCCA
	fr	0.291	0.303	**0.305**
	de	0.297	0.312	**0.318**
en	ar	0.283	0.295	**0.327**
	ru	0.281	0.301	**0.317**
	zh	0.283	0.303	**0.322**
avg.		0.287	0.303	**0.318**

表中展示了三种模型在 SimLex-999 语料上基于 Spearman 相关系数计算的单语词相似度评估值，值越高代表模型性能越好。L_1, L_2 代表不同的语言对。最好的结果用黑体标出，P-value 值 (p<0.15) 的结果用下划线标出，表示有明显的提升。

样展示上述三种模型的结果，并选用单语词向量的表现作为 base-line。如果两个模型的结果按照 Steiger 的方法计算得到的 P-value 值 (P<0.15)，我们则宣称取得了有效提升。

表 5-2　词向量的 QVEC-CCA 评估结果

L_1	L_2	Mono	BiCCA	BiKCCA
en	fr	**0.391**	0.377	0.382
	de	**0.391**	0.373	0.377
	ar	**0.391**	0.369	0.391
	ru	**0.391**	0.374	0.388
	zh	0.391	0.371	**0.402**
avg.		**0.391**	0.373	0.388

表中展示了三种模型在 SimLex-999 语料上的 QVEC-CCA 评估值，值越高代表模型性能越好。L_1, L_2 代表不同的语言对。最好的结果用黑体标出，BiCCA 和 BiKCCA 性能对比中 P-value 值 (p<0.15) 的结果用下划线标出，表示有明显的提升。

通过分析表中的结果，可以发现 BiKCCA 和 BiCCA 实验结果的总体趋势与前面的评估结果相同。不过有趣的一点是：无论是 BiCCA 还是 BiKCCA，QVEC-CCA 的得分均不如英语的单语词向量，一个比较合理的解释是 QVEC 是通过含有英语语言性质的矩阵来衡量单语词向量的质量，单语词向量在训练时已经综合了语言本身的性质。而对于 BiCCA、BiKCCA，由于训练得到的信息只是来自双语字词典，在单语基础上进行变化时单语的语言相关性质会被破坏，因此 QVEC-CCA 得分反而会不如单语词向量。

5.3.2 跨语言词典生成

表 5-3　跨语言词典生成结果

L_1	L_2	BiCCA			BiKCCA		
		top-1	top-10	top-50	top-1	top-10	top-50
en	fr	52.4	70.3	79.8	**53.3**	**71.1**	**79.9**
	de	53.2	72.4	80.4	**54.6**	**72.9**	**80.6**
	ar	32.7	53.1	62.1	**49.8**	**66.1**	**72.8**
	ru	37.3	55.4	63.4	**49.9**	**66.6**	**74.4**
	zh	36.3	60.1	73.1	**46.1**	**66.7**	**76.8**
avg.		42.4	62.3	71.8	**50.8**	**68.7**	**76.9**

表中展示了跨语言词典生成的准确度 (top-k accuracy, $k \in \{1, 10, 50\}$). 黑体的数字代表最好的结果。L_1, L_2 代表不同的语言对。当我们计算 MRR (mean reciprocal rank) 时，也会发现类似的趋势。

在表5-3中，我们展示了跨语言词典生成 (Cross-lingual Dictionary Induction) 任务 (§4.2) 的结果。表中的值是 $Top-k(k \in \{1, 10, 50\})$ 准确度。对于 gold 词典中的每一个词对 (e, f)，我们根据词向量在空间中的位置判断 f 是否在离 e 最近的 k 个邻居中。

从表5-3中我们可以发现：所有 BiKCCA 的结果均要比对应的 BiCCA 的结果高，无论是 $top-1, top-10, top-50$ 都有提升。这意味着通过捕获 BiCCA 所捕捉不到的非线性关系，BiKCCA 取得了性能上的提升，生成的双语词向量也有着更高的语义质量。这个结果也符合我们的假设 — 相对于线性模型，非线性模型能更好的描述语言间的关系。

同样地，对于本质上差异比较大的语言，如英语-俄语、英语-阿拉伯语、英语-汉语，BiKCCA 相比于 BiCCA 有着更显著的提升，平均取得了近 10 个百分点准确度的提升。

5.3.3 跨语言文本分类

在表5-4中，我们展示了不同模型在不同语言对上在跨语言文本分类 (Cross-lingual Document Classification) 任务 (§4.3) 的结果。表中主要数据代表 15 个话题上分类的 F1 值 (F1 值可以看成是准确率 (precision) 和召回率 (recall) 的加权平均值)。表中的第 4 行和第 8 行显示的是 Hermann and Blunsom 所训练的 MT system 的结果。值得注意的是，我们仅是将它列在这里作为参照，但并不期待我们的模型可以击败这个模型，因为这个系统相比于我们的系统在训练时拥有更多的语义信息。

表 5-4　跨语言文本分类结果

Setting	Languages				
[1pt]	French	German	Arabic	Russian	Chinese
En-L_2					
MT Baseline	**0.526**	**0.465**	**0.429**	**0.432**	-
BiCCA	0.446	0.399	0.344	0.276	0.204
BiKCCA	0.446	0.410	0.345	0.285	0.223
L_2-En					
MT Baseline	0.358	**0.469**	**0.448**	**0.404**	-
BiCCA	0.452	0.387	0.373	0.329	0.374
BiKCCA	**0.472**	0.435	0.387	0.346	0.375

表中展示了不同模型 TED 语料上跨语言文本分类任务的平均 F1-值，我们分别展示了两个方向的结果 ($En-L_2$：英语上训练，L_2 语言上测试，反之亦然)。黑色代表所有模型中最好的结果，下划线表示双语模型中最好的结果，我们引用了 Hermann and Blunsom 所训练的 MT system 作为 baseline。

通过比较线性模型 (BiCCA) 和非线性模型 (BiKCCA) 的结果，我们不难发现，几乎所有的 BiKCCA 的结果均好于 BiCCA，无论是英语-L_2 还是 L_2-英语两个方向。BiKCCA 所多捕捉的非线性因素对分类的准确度有着明显的作用。这也就说明：非线性模型生成的词向量相比于线性模型更适合来做跨语言模型迁移任务。

第 6 章 实验分析

本章中我们将对上章中得到的实验的结果、实验的误差、以及对我们提出的模型进行系统的分析，让读者对模型的理解得到提升。

6.1 实验结果分析

为了进一步解释为何 BiKCCA 能够在我们的评估任务中相较于 BiCCA 取得更好的结果，我们打算观察词向量的数学本质 — 词向量在向量空间的分布，进而对 BiKCCA 的性能提升作出解释。

图 6-1中给出了对词向量使用 t-SNE (Maaten and Hinton 2008[32]) 的方法得到的可视化结果。t-SNE 可视化方法与 PCA 可视化方法不同，PCA 是选取两个主成分维度进行显示，t-SNE 则是基于距离进行降维，t-SNE 可以保证原来在高维空间中比较近的两个点降到低维空间后仍然具有很短的距离。本图中我们分别展示了初始语言词向量以及将其进行线性变换 (BiCCA) 和非线性变换 (BiKCCA) 得到的结果。英语和汉语的词汇分别用红色和蓝色显示，单语和双语词向量的分布分别在左右两侧显示。对于每一对互译的词，我们的目标是分别为两个源语言学习各一个变化矩阵进而使这两个词拥有相似的表示，这在向量空间中体现为两个词向量的距离非常小。

通过对比图 6-1中 BiCCA 和 BiKCCA 的结果，我们有如下两个发现：

- 图中有些词对，如 (heaven, 天堂), (merchant, 商人)，BiKCCA 捕捉到了这些词对，而 BiCCA 没有捕捉到这些词对 (在 BiCCA 向量空间中，这些词的距离超出了捕获范围)
- 图中的词对，如 (notice, 注意), (consecutive, 连续)，这些词对都被两个模型成功捕捉，但 BiKCCA 空间中两个词的距离比 BiCCA 空间中的距离更近。

为了进一步说明第二个发现，我们计算了两个空间中每对词对应词向量的余弦相似度 (cosine similarity)[①]，并展示在表6-1中。

通过对比表 6-1中两个模型的结果，不难发现 BiKCCA 中几乎所有的词对的余弦相似度都大于 BiCCA 的结果，这也验证了图 6-1的第二个观察结果。

① 在计算之前，每个词向量均已经被单位化。

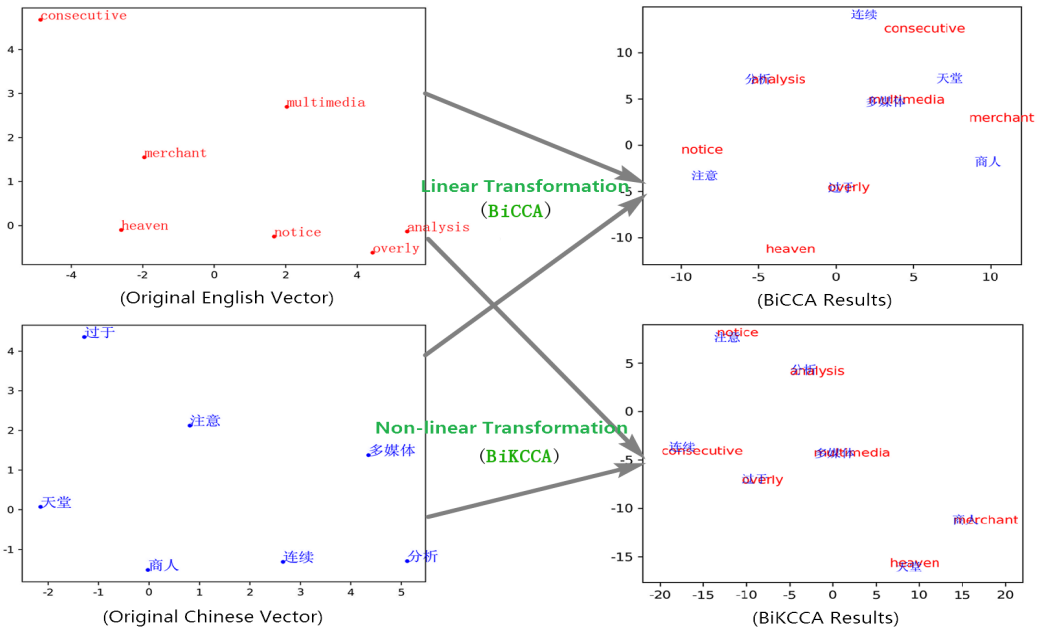

图 6-1　中英语料上一些高频词对应词向量的 t-SNE 可视化。英文和中文词分别用红色和蓝色字体标出。

表 6-1　图 6-1中部分词间的余弦相似度

word pair	Linear Trans (BiCCA)	Non-linear Trans (BiKCCA)
(analysis, 分析)	0.706	0.883
(multimedia, 多媒体)	0.863	0.935
(consecutive, 连续)	0.685	0.821
(heaven，天堂)	0.656	0.830
(merchant, 商人)	0.622	0.821

综合对图 6-1和表 6-1的观察，我们不难得出结论：互译的词在 BiKCCA 空间中比 BiCCA 更近，这使 BiKCCA 可以捕捉到 BiCCA 捕捉不到的一些词对。这也解释了 BiKCCA 在跨语言词典生成和跨语言文本分类两个任务中性能好于 BiCCA 的原因。

对于以上现象发生的原因，我们认为可以解释如下：语言间的关系既包括线性关系也包括非线性关系，原模型 (BiCCA) 只捕捉了语言间的线性关系，然后根据捕捉到的线性关系训练出一个对线性变换，进而生成词向量；可想而知，这个变换是不够充分的，它忽略了语言间的非线性关系，因此生成的不同语言词向量间的距离也不够理想。我们提出的 BiKCCA 模型则克服了此线性缺陷，将语言间的关系捕获的更充分，因此训练出的变换矩阵也更有效，生成的词向量也更能反映原词汇间本身的关系。

在图 6-2中，我们给出了三组英语中的近义词在空间中的分布情况 (词向量训练自中-英语料)。对于语义相近的词，我们希望在向量空间中能有相似的表示。

通过对比图中 BiCCA 和 BiKCCA 的结果，我们可以发现：

- 红色和蓝色的两组近义词在 BiKCCA 中的分布相比于 BiCCA 更加紧凑。
- BiCCA 和 BiKCCA 都能对反义词做到区分 (红蓝两个反义词组界限明显)。

第一个现象表明了 BiKCCA 产生的双语词向量相比于 BiCCA 在英语的单语语义上更加符合人们的期望，这也就解释了为什么 BiKCCA 能在英语单语词相似度评估任务上取得更高的成绩。

对于第一个现象出现的原因，我们的解释是：从上面对双语词向量的分析中，

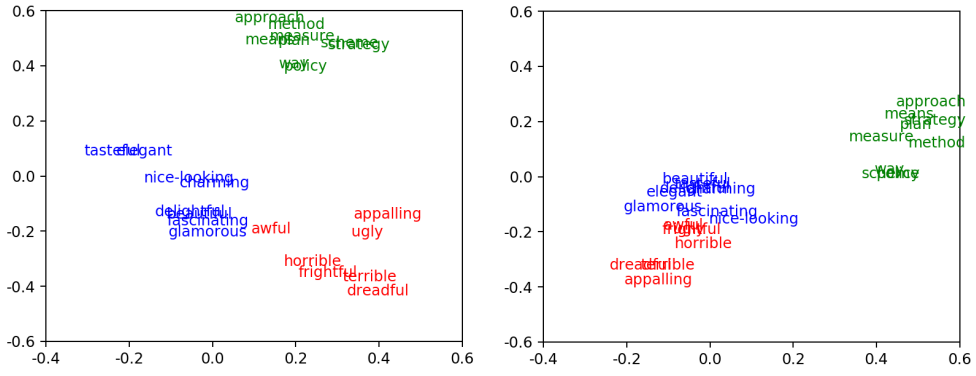

图 6-2　三组英语中的近义词在空间中的分布情况，不同组采用不同的颜色表示，左边代表 BiCCA 的结果，右边代表 BiKCCA 的结果。

我们可以得知，BiCCA 训练得到的一对转换矩阵是仅包含了语言间的线性关系，是不完善的，因此使用此矩阵对于英语单侧完成的转换也是不理想的；而 BiKCCA 训练出的转换矩阵融合了线性和非线性关系，理论上更加优秀，自然就有着更好的效果。

第二个现象的原因要归功于我们的训练数据，我们仅仅使用双语词典进行训练，而未涉及到单词的语境信息，而这本身是有助于区分反义词的。

最后，综合两个模型在以上三个评估实验中的表现，我们发现：对于本身关系比较密切的语言 (英-法，英-德)，BiKCCA 相比于 BiCCA 尽管在性能上有提升，但是提升不是很明显 (仅有 1 个百分点左右的提升)；而对于本身差异很大的语言如英-阿，英-俄，英-汉，模型获得的提升非常明显。假设对于所有语言对，我们已经捕获了几乎所有的非线性关系[①]，从定量的角度，我们可以初步得到如下结论：本身关系比较密切的语言之间大多是线性关系，非线性关系很少；而本身差异比较大的语言间除了线性关系以外还存在很大数量的非线性关系。

6.2 实验误差分析

我们找出了两个模型 (BiCCA 和 BiKCCA) 在跨语言词典生成任务中出错频率最高的一些词，并对这些词作了统计分析。发现一个非常典型的错误就是一个词经常被误翻译成一个和它有着相似但不同意义的词，比如词 “way/方法’’ 就容易被误翻译成 ‘‘政策/policy’’。这种错误在两种模型中都有出现，在 BiCCA 中出现很多，在 BiKCCA 中相对较少，但仍有一定数量。

对于这种误差的成因，解释如下：通过图 6-2，我们知道同一语言中，语义相近的词在向量空间中的分布也是相近的，比如"way", "policy" 这些词在英语空间中是聚集在同一个组中的，同理 ‘‘方法’’，‘‘政策’’ 在汉语空间中的分布也很相近。然而根据图 6-1，我们知道不同语言中互译的词的分布也是相近的。上述两点导致不同语言间所有语义相近的词全都 ‘‘挤’’ 在同一个小空间中，也就致使误翻译的情况出现很多。理论上这个问题是能够解决的，只要互译词间的距离足够近，趋于无穷小，那么这些 ‘‘噪声’’ 词则不能起到作用。我们提出的 BiKCCA 模型成功避免了部分错误，因为我们的转换矩阵相对比较理想，使得互译的词间的距离很近。而由于线性假设本身的缺陷，BiCCA 模型中这种错误出现的很多。

① 在测试的每个语言对上我们的第一相关系数已经到达 0.99。

6.3 模型分析

我们提出了一个简单易行，计算复杂度低而且容易扩展到大规模数据上的模型，并且在给定的 3 个评估任务中有着比 CCA-based 模型更好的性能。但我们的模型在理论上仍然存在以下两点缺陷：（1）我们只能为每个词学习一个词向量，而不是根据不同的词义学习多个词向量。事实上，同义词和多义词在自然语言中是非常常见的，而它们则经常是这类词向量生成算法的错误源泉。（2）我们现在只停留在词的级别上学习词向量，对于词的组合语义信息，我们是学习不到的。

第 7 章 相关工作

7.1 分布式表示 (Distributed Representations)

近年来有很多方法被提出用来学习语言的分布式表示。一种最简单的形式中，我们使用大型语料库中的分布式信息来学习词向量，方法是设定一个窗口，使用窗口中出现的词来计算当前词的词向量。这主要与主题建模 (topic-modelling) 有关，如 LSA（Dumais et.al，1988），LSI 和 LDA（Blei et.al，2003）等，这些模型使用文档级的上下文，并且倾向于捕获指定单词的主题而不是其更直接的句法语境。

另一类用来学习分布式表示的模型是神经语言模型 (Bengio et al., 2003)，近年来神经网络模型在分布式表示学习领域已经受到很多关注并且产生了令人称赞的结果 (Collobert and Weston,2008; Mnih and Hinton, 2009; Mikolov et al., 2010)。Collobert et al. (2011) 进一步提出了使用神经网络结构在大规模平行语料上学习词向量，并说明训练出的词向量可以用来改进标准的监督任务。

无监督的词的表示 (词向量) 可以很容易插入到 NLP 各种相关任务中，并在这些上这些任务取得了成功，如主题建模 (Blei et al., 2003)，命名实体识别 (Turian et al., 2010; Collobert et al., 2011)。

7.2 跨语言表示的学习

大多数对分布式表示的研究都是专注于单一语言。其中英语，靠着它大量的资源、语言的广泛度收获了大批学者的注意力。然而由于现实需求，其他各个语言的工作也并一刻没有停留，依靠各种类型的平行语料，人们提出了各种各样的模型来进行跨语言学习，将所有语言的词向量根据语义表示在一个共享的向量空间中。

与我们任务相关的是，Yih et al. (2011) 提出了 S2Nets 来在可比语料上为 tf-idf 学习词向量，他们的模型优化文档的余弦相似度，使用学习过程中的语义相似度分数进行评估。Sarath Chandar et al. (2013) 训练了一个跨语言编码器，使用自动编码器在两个语言的平行语料中重新编码单词，这是 Ngiam et al. (2011) 提出的模型的加强版本。Hermann and Blunsom (2014) 提出了一个在句对齐语料上使用 CVM 来学习多语词向量的方法，并在跨语言文本分类任务上取得了当时最好的结

果。Mikolov et al. (2013a), Mikolov et al. (2013b) 提出仅使用词典作为跨域言知识来学习一个变换矩阵从而将一个语言的词向量转移到另一个语言的方法，这也为后来的很多工作提供了方向。

结　论

词向量作为当前各种自然语言处理任务中不可或缺的因素，一直在被广泛的研究，跨语言词向量作为普通单语词向量的扩充，不仅提高了单语词向量的质量，也为许多跨语言任务带来了新鲜血液。近年来已经有很多模型被提出来训练跨语言词向量，我们对以往的模型进行了系统考察，并对前人的理论加以改进，提出了我们自己的模型。

我们的主要成果和贡献如下：

（1）我们提出了一种基于 CCA 的方法来学习跨语言的词向量，它克服了原模型[①]的缺点。

（2）通过在三个任务、五种语言对上的综合评估，我们展示了新模型比原模型有着更好的性能，尤其是在本身差异比较大的语言对上产生了很好的效果 (有的任务上获得了近 10 个百分点的准确度提升)。

（3）我们定量的分析了不同语言对间的关系，发现：本身关系比较密切的语言之间大多是线性关系，非线性关系很少；而本身差异比较大的语言间除了线性关系以外还存在很大数量的非线性关系。我们相信这个发现可以给以后的研究提供方向。

当然，我们的模型本身也存在不足，具体细节已经在文章中指出，将来我们会：

（1）根据理论上的不足，进一步考察和改进模型的性能。

（2）考察更多的语言对，进一步验证我们的发现。

（3）将我们的模型应用到更多的下游任务中去，如跨语言依存分析，机器翻译等。

① 这里的原模型指的是 Faruqui 和 Dyer 提出的 BiCCA 模型

参考文献

[1] Socher R, Bauer J, Manning C D, et al. Parsing with Compositional Vector Grammars[C] // . Sofia, Bulgaria : Association for Computational Linguistics, 2013 : 455 – 465.

[2] Dyer C, Ballesteros M, Ling W, et al. Transition-Based Dependency Parsing with Stack Long Short-Term Memory[C] // . Beijing, China : Association for Computational Linguistics, 2015 : 334 – 343.

[3] Guo J, Che W, Yarowsky D, et al. Cross-lingual Dependency Parsing Based on Distributed Representations[C] // . Beijing, China : Association for Computational Linguistics, 2015 : 1234 – 1244.

[4] Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind[C] // . 2015 : 151 – 159.

[5] Faruqui M, Dyer C. Improving Vector Space Word Representations Using Multilingual Correlation[C] // . Gothenburg, Sweden : Association for Computational Linguistics, 2014 : 462 – 471.

[6] Uszkoreit J, Ponte J M, Popat A C, et al. Large Scale Parallel Document Mining for Machine Translation[C] // . Stroudsburg, PA, USA : Association for Computational Linguistics, 2010 : 1101 – 1109.

[7] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities among Languages for Machine Translation[J]. Computer Science, 2013.

[8] Zhang J, Liu S, Li M, et al. Bilingually-constrained Phrase Embeddings for Machine Translation[C] // . Baltimore, Maryland : Association for Computational Linguistics, 2014 : 111 – 121.

[9] Wan X. Co-training for Cross-lingual Sentiment Classification[C] // . Stroudsburg, PA, USA : Association for Computational Linguistics, 2009 : 235 – 243.

[10] Klementiev A, Titov I, Bhattarai B. Inducing Crosslingual Distributed Representations of Words[C] // . 2012.

[11] Shi T, Liu Z, Liu Y, et al. Learning Cross-lingual Word Embeddings via Matrix Co-factorization[C] // . Beijing, China : Association for Computational Linguistics, 2015 : 567 – 572.

[12] Upadhyay S, Faruqui M, Dyer C, et al. Cross-lingual Models of Word Embeddings: An Empirical Comparison[C] // . Berlin, Germany : Association for Computational Linguistics, 2016 : 1661 – 1670.

[13] McDonald R, Nivre J, Quirmbach-Brundage Y, et al. Universal Dependency Annotation for Multilingual Parsing[C] // . Sofia, Bulgaria : Association for Computational Linguistics, 2013 : 92 – 97.

[14] Gouws S, Bengio Y, Corrado G. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments[C] // Blei D, Bach F. . [S.l.] : JMLR Workshop and Conference Proceedings, 2015 : 748 – 756.

[15] Hermann K M, Blunsom P. Multilingual Models for Compositional Distributed Semantics[C] // . Baltimore, Maryland : Association for Computational Linguistics, 2014 : 58 – 68.

[16] Akaho S. A kernel method for canonical correlation analysis[J]. CoRR, 2006, abs/cs/0609071.

[17] Hotelling H. Relations Between Two Sets of Variates[J]. Biometrika, 1936, 28(3/4) : 321 – 377.

[18] Fukumizu K, Bach F R, Gretton A. Statistical Consistency of Kernel Canonical Correlation Analysis[J]. J. Mach. Learn. Res., 2007, 8 : 361 – 383.

[19] Myers J L, Well A A. Research design and statistical analysis[M]. 1nd ed. [S.l.] : Mahwah, N.J. : Lawrence Erlbaum Associates, 1995.

[20] Hill F, Reichart R, Korhonen A. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation[J]. CoRR, 2014, abs/1408.3456.

[21] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing Search in Context: The Concept Revisited[C] // . New York, NY, USA : ACM, 2001 : 406 – 414.

[22] Ammar W, Mulcaire G, Tsvetkov Y, et al. Massively Multilingual Word Embeddings[J]. CoRR, 2016, abs/1602.01925.

[23] Steiger J H. Tests for comparing elements of a correlation matrix.[J]. Psychological Bulletin, 1980, 87(2) : 245.

[24] Vulić I, Moens M-F. Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses[C] //. Atlanta, Georgia : Association for Computational Linguistics, 2013 : 106 – 116.

[25] Zhang M, Liu Y, Luan H, et al. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation[C] //. [S.l.] : AAAI Press, 2016 : 2870 – 2876.

[26] Bond F, Foster R. Linking and Extending an Open Multilingual Wordnet[C] //. Sofia, Bulgaria : Association for Computational Linguistics, 2013 : 1352 – 1362.

[27] Lewis D D, Yang Y, Rose T G, et al. RCV1: A New Benchmark Collection for Text Categorization Research[J]. J. Mach. Learn. Res., 2004, 5 : 361 – 397.

[28] Cettolo M, Girardi C, Federico M. WIT^3: Web Inventory of Transcribed and Translated Talks[C] //. 2012 : 261 – 268.

[29] Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval[J]. Inf. Process. Manage., 1988, 24(5) : 513 – 523.

[30] Freund Y, Schapire R E. Large Margin Classification Using the Perceptron Algorithm[J]. Machine Learning, 1999, 37(3) : 277 – 296.

[31] Dyer C, Lopez A, Ganitkevitch J, et al. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models[C] //. Uppsala, Sweden : Association for Computational Linguistics, 2010 : 7 – 12.

[32] Maaten L v d, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(Nov) : 2579 – 2605.

[33] Hardoon D R. Semantic models for machine learning[D]. [S.l.] : University of Southampton, 2006.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《基于 CCA 的跨语言语义表示的研究与实现》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：　　日期：　　年　　月　　日

致　谢

衷心感谢导师 曹海龙 老师对本人的精心指导。从最初的课题确定，到语料准备，再到实验研究与论文的撰写，曹海龙老师给了我耐心的指导和无私的帮助，通过这几个月的努力，我受益良多，不仅掌握了很多机器翻译相关的知识，还能在实践之中加以运用，这与曹老师的指导是分不开的，他的言传身教将使我终生受益。

感谢 赵铁军 教授，以及实验室全体老师和同窗们的热情帮助和支持！

附录 1 带正则项的 KCCA 求解过程

我们给出了 Hardoon[33] 在 Semantic Models for Machine Learning①一文中对正则化 KCCA 的定义和求解过程。

1.1 KCCA with regularisation

In the following section we provide an alternative solution to the regularisation of the kernel CCA optimisation problem. Due to the fact that we compute the eigenproblem for both views combined. We regularise kernel CCA by adding the weight term to the constraint of the optimisation problem such that $max_{\alpha,\beta}\rho = \alpha' K_a K_b \beta$ is now subject to:

$$(1-\tau)\alpha' K_a^2 \alpha + \tau \alpha' K_a \alpha = 1 \tag{1-1}$$

$$(1-\tau)\beta' K_a^2 \beta + \tau \beta' K_a \beta = 1 \tag{1-2}$$

1.2 Mathematical solution

The corresponding Lagrangian to KCCA Optimisation problem is:

$$\begin{aligned} L(\lambda_a, \lambda_b, \alpha, \beta) = \alpha^T K_a K_b \beta - \frac{\lambda_a}{2}((1-\tau)\alpha' K_a^2 \alpha + \tau \alpha' K_a \alpha - 1) \\ -\frac{\lambda_b}{2}((1-\tau)\beta' K_a^2 \beta + \tau \beta' K_a \beta - 1) \end{aligned} \tag{1-3}$$

Taking derivatives in respect to α and β, we obtain:

$$\begin{aligned} \frac{\partial L}{\partial \alpha} &= K_a K_b \beta \lambda_a ((1\tau) K_a^2 \alpha + \tau K_a \alpha) \\ &= K_a K_b \beta \lambda_a K_a ((1\tau) K_a + \tau I)\alpha = 0 \end{aligned} \tag{1-4}$$

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= K_b K_a \alpha \lambda_b ((1\tau) K_b^2 \beta + \tau K_b \beta) \\ &= K_b K_a \alpha \lambda_b K_b ((1\tau) K_b + \tau I)\beta = 0 \end{aligned} \tag{1-5}$$

Following the same procedure as in Section 2.2 we are able to find that $\lambda_a = \lambda_b$, let $\lambda = \lambda_a = \lambda_b$. Consider the case where the kernel matrices are invertible, we have:

① https://eprints.soton.ac.uk/262019/

$$\begin{aligned}\beta &= \frac{((1\tau)K_b + \tau I)^{-1}K_b^{-1}K_bK_a\alpha}{\lambda} \\ &= \frac{((1\tau)K_b + \tau I)^{-1}K_a\alpha}{\lambda}\end{aligned} \tag{1-6}$$

substituting into equation 1-4 gives:

$$K_aK_b((1\tau)K_b + \tau I)^{-1}K_a\alpha = \lambda^2K_a((1\tau)K_a + \tau I)\alpha \tag{1-7}$$

multiplying both sides of the equation by K_1^a gives:

$$K_b((1\tau)K_b + \tau I)^{-1}K_a\alpha = \lambda^2((1\tau)K_a + \tau I)\alpha. \tag{1-8}$$

Observe that the left component of the generalised eigenproblem is non symmetric, resulting in non orthogonal eigenvectors. We can use partial Gram-Shmidt orthonormalisation in order to obtain an eigenproblem with symmetric matrices.