哈爾濱工業大學

毕业设计（论文）开题报告

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　2017 年 3 月 15 日

哈尔滨工业大学教务处制

说　　明

一、开题报告主要内容

1．课题来源及研究的目的和意义

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，融入多语言学知识源信息的词嵌入表示和基于神经网络的词嵌入表示开始应用在许多跨语言自然语言处理的任务中并且表现出了优异的性能。近几年来更是出现了各种各样的用于词嵌入的跨语言的模型，这些模型大多取得了不错的成果，但是他们训练所用到的跨语言语料不尽相同，有的用到的是文档级对齐语料，有的用到的是句子级的对齐，还有的用到的是词对齐语料。我们都知道使用高代价的语料（如词对齐）往往能取得令人满意的成果，但是对于具体的任务来说，权衡利弊后，使用低代价语料的模型似乎能在满足准确率的条件下最大限度地减少项目花费。

经过我们的考察，发现基于 CCA 的跨语言学习模型在理论上还有模型复杂度均有着还不错的成果，所以我们打算先对现存的这种模型进行仔细地考察，定性和定量的考察此种模型的优缺点，然后针对它的缺点进行优化和改进，争取实现一个有着更好性能的模型。

2．国内外在该方向的研究现状及分析

通过词的分布信息利用单语语料训练词向量在 NLP 领域已经是普遍存在的现象,而现在关于跨语言模型的猜想已经得到验证：

1.词向量的质量可以通过加入跨语言的元素得到提升(Klementiev et al., 2012; Zou et al., 2013; Vuli´ c andMoens,2013b; Mikolovetal.,2013b; Faruqui and Dyer, 2014; Hermann and Blunsom, 2014; Chandar et al., 2014, inter alia）

2.跨语言元素的加入使词向量对于单语（Faruquiand Dyer,2014;Rastogietal.,2015)和双语(Guo et al., 2015; Søgaard et al., 2015; Guo et al.2016) 的任务完成的更好。

在近几年通过跨语言模型来生成词向量的工作中，下面几个模型取得了较好的成绩并且代表了几种典型的方向：

- 利用文档级对齐训练语料的模型[1] (Vuli´c and Moens, 2015)
- 利用句子级对齐训练语料的模型[2](Hermann and Blunsom, 2014)(Gouws et al.2015)
- 利用词级对齐训练语料的模型[3](Faruqui and Dyer, 2014; Gouws et al.2015)
- 利用词级对齐和句子级对齐训练语料的模型[4](Luong et al.2015)

高代价的语料（如词对齐）往往能取得令人满意的成果，但是对于具体的任务来说，花费少而精度也能达标的低代价模型同样具有吸引力；如此看来，设计一个复杂度低而精度相对比较高的模型工作是十分有意义的。

3．主要研究内容

3.1　本课题的主要研究内容分为以下两部分：

1. 对以下提到的模型本身和通过使用这个模型所训练出来的词向量去完成设计好的分别针对单语和双语的词语相似度、语义和句法任务，分别定性和定量评估此模型的特点。

2. 通过完成第一部分的考察和实验后对原模型进行改进，争取设计出一个性能更好的模型。

3.2　本课题涉及到的主要模型如下(括号里表示简称，之后的内容多会使用这里提到的简称表示该模型)

- Canonical-Correlation Analysis(CCA) (T. R. Knapp)
- Bilingual Correlation Based Embeddings(BiCCA)

3.3 Canonical-Correlation Analysis(CCA) (T. R. Knapp)

3.3.1　CCA 原理分析

典型关联分析（CCA）是利用综合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。举个简单的例子，我们想考察一个人解题能力 X（解题速度$x1$， 解题正确率$x2$）与他/她的阅读能力 Y（阅读速度$y1$，理解程度$y2$）之间的关系；

CCA 的做法就是寻找两个投影向量 a，b,分别与 X,Y 相乘得到

$$X' = a^T X \ , Y' = \ b^T Y \tag{1}$$

然后使用 X’ 与 Y’ 的 Pearson 相关系数

$$\rho(X',Y') = \frac{cov(X',Y')}{\sqrt{E[X'^2]E[Y'^2]}} = \frac{E[(X-u_x)(Y-u_y)]}{\sqrt{E[X'^2]E[Y'^2]}} \tag{2}$$

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$\rho(X',Y')$最大，这样得到的 a 和 b 就是使得 X 和 Y 就有最大关联的权重,可以简单描述如下：　　$a,b = \arg\max_{a,b} \rho(X',Y')$,其中$X',Y'$来自于(1)　　(3)

3.3.2　CCA 的求解过程：

给定两组向量 X 和 Y，X 维度为 p1， Y 维度为 p2，默认 p1 ≤ p2。形式化表示如下

$$T = \begin{bmatrix} X \\ Y \end{bmatrix},\ E(T) = \begin{bmatrix} u1 \\ u2 \end{bmatrix}, \Sigma = Var(T) = \begin{bmatrix} \Sigma 11 & \Sigma 12 \\ \Sigma 21 & \Sigma 22 \end{bmatrix} \tag{4}$$

Σ是 T 的协方差矩阵；左上角是 X 自己的协方差矩阵；右上角是 Cov(X,Y)；左下角是 Cov(Y,X)，也是Σ12 的转置；右下角是 Y 的协方差矩阵

与之前一样，我们从 X 和 Y 的整体入手，定义 $X' = a^T X \quad Y' = \ b^T Y$ 我们可以算出 X' 和 Y' 的方差和协方差：

$$Var(X') = a^T \Sigma 11 a \ , Var(Y') = b^T \Sigma 22 b \ , Cov(X',\ Y') = a^T \Sigma 12 b \tag{5}$$

其中

$$Var(X') \ = \ Var(a^T X) = \frac{1}{N}\sum_{i=1}^{n}(a^T X - a^T u1)^2 = \ a^T \Sigma 11 a \tag{6}$$

同理可推得$Var(Y')$, $Cov(X',\ Y')$

由公式（2）（5）可得：

$$\rho(X',Y') \ = \ \frac{a^T \Sigma 12 b}{\sqrt{a^T \Sigma 11 a \cdot b^T \Sigma 22 b}} \tag{7}$$

现在问题就演化成：

$$\text{Maximize } a^T \Sigma 12 b$$
$$\text{Subject to: } a^T \Sigma 11 a = 1,\ b^T \Sigma 22 b = 1 \tag{8}$$

构造 Lagrangian 等式:

$$\mathrm{L} = a^T \Sigma 12\mathrm{b} - \frac{\lambda}{2}\left(a^T \Sigma 11\mathrm{a} - 1\right) - \frac{\theta}{2}(b^T \Sigma 22\mathrm{b} - 1) \quad (9)$$

求导，化简后得：

$$\Sigma 12b - \lambda \Sigma 11a = 0 \quad (10)$$

$$\Sigma 21a - \theta \Sigma 22b = 0 \quad (11)$$

由公式(8),(10),(11)可得

$$\lambda = \theta = a^T \Sigma 12b \quad (12)$$

所以只需要求最大的λ即可。

将公式（11）待入（10）中，利用$\lambda = \theta$的条件，可以得到下面的式子：

$$\Sigma 11^{-1} \Sigma 12 \Sigma 22^{-1} \Sigma 21\mathrm{a} = \lambda^2 \mathrm{a} \quad (13)$$

这样先对矩阵$\Sigma 11^{-1} \Sigma 12 \Sigma 22^{-1} \Sigma 21$求特征值$\lambda^2$和特征向量$a$，然后根据公式（10）求得 b；

由此，得到了λ最大时的$a1$和$b1$。那么$a1$和$b1$称为典型变量（canonical-variates），λ即是X'和Y'的 pearson 相关系数。

除了以上介绍的典型变量外，还有 rank(A)-1 对投影向量。首先，在求出第一对典型变量的基础上求第二对典型变量。由上述分析我们可以知道该模型为：

$$\begin{aligned} &\text{Maximize } {a_2}^T \Sigma 12 b_2 \\ &\text{Subject to: } {a_2}^T \Sigma 11 a_2 = 1,\ {b_2}^T \Sigma 22 b_2 = 1 \\ &\qquad\qquad {a_1}^T \Sigma 11 a_2 = 0,\ {b_1}^T \Sigma 22 b_2 = 0 \end{aligned} \quad (14)$$

类似于公式(9)，构造 Lagrangian 等式:

$$\mathrm{L} = {a_2}^T \Sigma 12 b_2 - \frac{\lambda}{2}\left({a_2}^T \Sigma 11 a_2 - 1\right) - \frac{\theta}{2}\left({b_2}^T \Sigma 22 b_2 - 1\right) + \gamma({a_1}^T \Sigma 11 a_2) + \delta({b_1}^T \Sigma 22 b_2) \quad (15)$$

令导数得 0，化简后可以得到：

$${a_2}^T \Sigma 12 b_2 - \lambda {a_2}^T \Sigma 11 a_2 = 0 \quad (16)$$

$${b_2}^T \Sigma 12 a_2 - \theta {b_2}^T \Sigma 11 b_2 = 0 \quad (17)$$

由公式(16)，(17)分别与(14)联立可得：

$$\Sigma 12 b_2 - \lambda \Sigma 11 a_2 = 0 \quad (18)$$

$$\Sigma 12 a_2 - \theta \Sigma 11 b_2 = 0 \quad (19)$$

由(18)，(19)可得

$$\lambda = \theta = a2^T \Sigma 12 b \quad (20)$$

$$\Sigma 11^{-1} \Sigma 12 \Sigma 22^{-1} \Sigma 21 a_2 = \lambda^2 a_2 \quad (21)$$

可以发现，公式(21)与之前推得的公式(13)完全一致，这说明a_2，b_2也都是矩阵$\Sigma 11^{-1} \Sigma 12 \Sigma 22^{-1} \Sigma 21$的特征向量，而$a_2$ b_2是不同于a_1，b_1的，因为条件${a_1}^T \Sigma 11 a_2 = 0$，${b_1}^T \Sigma 22 b_2 = 0$的约束，因此，可以推得一个重要结论：

$\Sigma 11^{-1} \Sigma 12 \Sigma 22^{-1} \Sigma 21$的所有特征值和特征向量的组合就是我们要求解的 rank(A)个相关系数和投影向量。

由此,CCA 要求解的所有内容都得到了，因此可以进行下一步对高等级模型的研究工作。

3.4 Bilingual Correlation Based Embeddings(BiCCA) [3]

3.4.1 BiCCA 原理简析：

BiCCA 的原理是先利用训练语料得到双语的 CCA 模型的投影矩阵，再对两种语言的单语词向量进行投影，将投影的得到的新的向量作为“增加了双语知识”的词向量作为输出。

3.4.2 BICCA 求解过程：

a)BiCCA 的输入是两个单语语料库的训练的词向量。

b)BiCCA 先人工挑选出 N 组可以互相翻译的词向量对（wi,vi），形成向量矩阵 W’ 和 V’，再根据W’ 和V’训练 CCA 模型：

$$P_w, P_v = CCA(\mathrm{W}', \mathrm{V}') \tag{22}$$

再令 $\mathrm{W}^* = WP_w$ ， $\mathrm{V}^* = VP_V$ (23)

由此就得到了“增加了双语知识”的词向量。

再利用得到的词向量去进行单语和双语的评估实验，进而评估词向量的质量。

4. 研究方案

4.1 第一部分：完成对现有模型的考察工作，主要分为以下三方面：

1. 对模型训练出的词向量进行单语和双语的评估
2. 对该模型进行定量的分析：
   a) 评估该跨语言模型相对单语模型的性能提升程度
   b) 高代价语料模型相对此模型的性能提升
3. 对该模型的表现进行定性分析：
   a) 跨语言模型之所以能提升性能的本质
   b) 对比该模型与其他模型的词向量在词空间的分布

4.1.1 第一部分的实验设计：

1. 单语评估：基于词相似度的评估方案
   a) 实验目的：评估该跨语言模型相对于单语模型是否在词相似度实验上有提升
   b) 实验评估方法：分别是 Spearman’s rank correlation coefficient[5] (Myers and Well, 1995)和 QVEC[6] (Tsvetkov et al.2015)
2. 跨语言字典生成：
   a) 实验目的：考察模型跨语言寻找近/同义词的能力
   b)实验评估方法：The task of cross-lingual dictionary induction[1] (Vuli´candMoens,2013a)
3. 跨语言文本分类：
   a) 实验目的：考察模型跨语言进行文本分类的能力
   b) 实验方案：CLDC [7](Klementiev.2012)

4.2 第二部分：

通过完成第一部分，进一步了解这个模型的优点与缺点，争取找到导致模型缺点的本质原因，对缺点进行改进与优化，设计出更好的模型。在保持输入与输出的形式不变的条件下提升该模型的准确度。

5．进度安排，预期达到的目标

5.1　进度安排：

第 1-3 周：	根据课题了解相关技术并考察相关模型，完成基本实验设计工作，参加开题答辩。
第 4-6 周：	基本完成设计的实验，并根据需要补充实验。
第 7-9 周	分析实验结果，进一步深入理解模型的优缺点，找出模型不够完美的原因。并初步提出解决方案，参加中期答辩。
第 10-12 周	进一步寻找解决方案，并跟据解决方案优化模型，通过查阅相关文章完成模型优化过程。
第 13-15 周	整合系统，参数微调，撰写结题答辩报告，参与结题答辩

5.2　预期达到的目标

1. 通过理论知识的进一步补充和实验结果分析在中期答辩前成功找到模型不够完美的原因，并初步提出解决方案
2. 中期答辩后进一步寻找解决方案，并成功根据方案优化模型。
3. 保质保量完成毕设内容，参与结题答辩。

6．课题已具备和所需的条件、经费

6.1　课题所需的条件，经费：

本课题需要前面实验所需的各种语料，具体包括：

1. 词向量训练用到的跨语语料：
   - Europarl v7 parallel 语料[8](Koehn,2005)，FBIS parallel(LDC2003E14)
   - UN-corpus 语料
2. 词相似实验用到的语料：
   - SimLex dataset for English[9] (Hill et al.2014)
3. 跨语言字典生成实验的语料:
   - Gold Dictionaries Derived From Open Multilingual WordNet data[10] (Bond and Foster 2013)

4. 跨语言文本分类实验语料：

• RCV2 Reuters multilingual corpus

本课题的实验环境已经具备，实验语料也不需要人为进行标注，因此不需要经费支持。

6.2 已具备条件：

1. 经过开题前的准备，已经找到了一部分实验所需的语料
2. 实验室已经提供了能训练模型并运行实验的服务器账号
3. 已经基本了解实验所用模型的原理

7．研究过程中可能遇到的困难和问题，解决的措施

1. 做实验时可能会遇到模型的 bug，测试结果与论文结果不符等问题；
   解决措施：难以解决的 bug 通过联系原作者咨询 bug 的解决方法。对于异常的测试结果，多进行几次实验，分析原因，如果确认是原论文的问题，会联系论文作者寻求帮助。
2. 理论知识不足：
   解决措施：积极补充相关理论知识，阅读前沿文章，了解当下比较流行的技术与理论。
3. 能找出现有模型的缺点，但难以给出合理的解决方案；
   解决措施：1.积极查看相关论文，了解前沿最新的进展，从论文中寻找启发点。
   2.向老师咨询，寻求指导。

8．主要参考文献

[1] Ivan Vuli´c and Marie-Francine Moens.2013a. Crosslingual semantic similarity of words as the similarity of their semantic word responses. In Proc. of NAACL.

[2] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics

[3] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word represent- tations using multilingual correlation. In Proc. of EACL

[4] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In Proc. of the Workshop on Vector Space Modeling for NLP.

[5] Jerome L. Myers and Arnold D. Well. 1995. Research Design & Statistical Analysis. Routledge.

[6] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. arXiv preprint arXiv:1602.01925

[7] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In Proc. of COLING.

[8] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In Proc. of MT Summit

[9] Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. arXiv preprint ar-Xiv:1408.3456.

[10] Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In Proc. of ACL.

[11] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In ACL.