文章编号：1003-0077（2011）00-0000-00

基于点关联测度矩阵分解的中英跨语言词嵌入*

于东[1,2]，赵艳[2]，韦林煊[2]，荀恩东[1,2]

（1.北京语言大学 大数据与教育技术研究所，北京 100083；

2.北京语言大学 信息科学学院，北京 100083）

摘要：研究基于矩阵分解的词嵌入方法，提出统一的描述模型，并应用于中英跨语言词嵌入问题。以双语对齐语料为知识源，提出跨语言关联词计算方法和两种点关联测度的计算方法：跨语言共现计数和跨语言点互信息。分别设计目标函数学习中英跨语言词嵌入。从目标函数、语料数据、向量维数等角度进行实验，结果表明：在中英跨语言文档分类中以前者作为点关联测度最高得到87.04%的准确率；在中英跨语言词义相似度计算中，后者作为点关联测度得到更好的性能，同时在英-英词义相似度计算中的性能略高于主流的英语词嵌入。

关键词：点关联测度；词嵌入；跨语言；矩阵分解

中图分类号：TP391　　**文献标识码：**A

Chinese-English Cross-lingual Word Embeddings Based on Pointwise Relevant Measurement Matrix Factorization

YU Dong[1,2], ZHAO Yan[2], WEI Linxuan[2] , XUN Endong[1,2]

(1. Institute of Big Data and Language Education, Beijing Language and Culture University, Beijing 10083, China; 2.College of Information Science, Beijing Language and Culture University, Beijing 10083, China)

Abstract: This paper presents a unified model for matrix factorization based word embedding, and applies the model to Chinese-English Cross-lingual word embedding. It proposes a method to determine Cross-lingual relevant word on parallel corpus. Both Cross-lingual word co-occurrence and pointwise mutual information are served as pointwise relevant measurements to design objective function for learning Cross-lingual word embeddings. Experiments are carried out from perspectives of different objective function, corpus, and vector dimension. For the problem of Cross-lingual document classification, the best performance model achieves 87.03% in accuracy, as it adopts Cross-lingual word co-occurrence as relevant measurement. However, models adopt Cross-lingual pointwise mutual information get better performance in Cross-lingual word similarity calculation task. Meanwhile, for the problem of English word similarity calculation, experimental result shows that our methods get slightly higher performance than English word embeddings trained by mainstream method.

Key words: pointwise relevant measurement; word embedding; Cross-lingual; matrix factorization

1 引言

词嵌入（Word Embeddings）可以将自然语言中的每个词表示为稠密、低维的连续实数向量，在基于深度神经网络的自然语言处理方法中，词嵌入往往作为预处理步骤，起到基础性作用，是目前语言信息处理领域研究和应用的热点问题。跨语言词嵌入（Cross-lingual Word Embeddings，CWE）是该领域的一个分支，目的是将两种甚至多种语言的词汇以向量形

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金资助项目(61300081)；国家高技术研究发展计划“863”计划项目(2015AA015409)；中央高校基本科研业务费专项资金资助项目(北京语言大学科研项目:16YJ030002)

式表示在同一个向量空间中，从而能够直接通过向量计算来描述不同语言词汇之间的关系，为跨语言信息处理提供良好表示形式，在近一段时间受到广泛关注。Klementiev[1]等人最早提出跨语言词嵌入问题。近几年许多学者针对该问题开展研究，第一类方法是采用新的学习框架学习跨语言映射知识，如基于自动编码器[2]、典型相关性分析[3]等。第二类方法通过语料变换和洗牌，将跨语言词向量问题转变为普通词向量问题[4][5]。在应用方面，跨语言词嵌入被应用于机器翻译[6]、双语词典抽取[4]、句法分析[7]等任务，均取得良好的效果。目前，跨语言词嵌入主要针对英语、德语等西方语言，相关研究在在国内开展较少，尚无针对中文的研究成果发表。

目前，基于矩阵分解的词嵌入学习逐渐受到重视，Pennington[8]和 Levy[9]均对此问题进行探讨。本文认为词嵌入可以由词语的点关联测度经矩阵分解学习得到，并给出统一的目标函数形式。在此基础上，将该方法扩展用于中英跨语言词嵌入。本文以中英对齐语料为主要知识源，提出跨语言关联词和点关联测度的计算方法，分别探讨以跨语言共现计数(Cross-lingual Co-occurrenc)和跨语言点互信息(Cross-lingual PMI)作为点关联测度时的目标函数，用于学习跨语言词嵌入表示，并以跨语言文档分类（Cross-lingual Document Classification,CDC）和跨语言语义相似度(Cross-lingual Similarity，CLS)评价词嵌入性能。实验中，通过对比不同目标函数、不同知识源、不同维度下跨语言词嵌入的性能，验证本文训练跨语言词嵌入的有效性，并从适用问题、应用领域等方面给出综合分析。

本文第 2 节介绍跨语言词向量训练的相关工作，第 3 节具体介绍模型和方法，第 4 节介绍跨语言词向量的应用问题，第 5 部分针对 CDC 和 CLS 等任务进行实验和分析，第 6 部分给出结论和未来工作。

2 相关工作

2.1 跨语言词嵌入相关研究

跨语言词嵌入问题由 Klementiev[1]等人提出，首先借助神经网络语言模型构建初始词向量，然后借鉴多任务学习框架，利用对齐语料的词共现特征导出跨语言词嵌入。此后，许多学者对该问题进行研究，提出不同的学习模型。Faruqui[3]、Zou[6]等工作将跨语言词嵌入分为两步，首先分别训练单语言词嵌入，然后以两者的某种距离作为目标函数，学习得到跨语言词嵌入。由于采用串行级联形式，该方法难以同时学习单语言和跨语言的嵌入表示。Hermann and Blunsom[10]、Chandar A P[2]等工作以对齐语料中的句子作为训练单元，通过组合词向量构成句向量，再以句子向量距离、两个语言作为目标函数学习词嵌入。这种方式对于句子级别的表达具有较高的性能，但缺乏对词之间的语义表达。在目标函数设计方面，Gouws[11]分别设计单语、跨语目标函数，然后累加得到总目标函数训练跨语言词嵌入，该思路也被 Soyer[12]、Shi[13]等工作采用。2015 年后 Gowus[5]、Vulic[6]、 Coulmance[14]等工作分别设计算法对齐语料进行随机词混合，将得到的混合语料作为训练数据，将跨语言词嵌入转换为单一语言词嵌入，也得到了较好的效果。

目前，跨语言词嵌入仍然是表示学习的一个研究热点问题，并开始逐渐向多语言、多粒度、多功能的方向发展，在跨语言文档分类、跨语言情感分类、跨语言相似度计算、机器翻译、跨语言句法分析等领域得到应用。

2.2 基于矩阵分解的词嵌入

自 2013 年 Mikolov[15]开源 word2vec 工具后，词嵌入的研究和应用逐渐形成热潮。2014 年 Levy[9]发表文章，证明 Mikolov 提出的基于负样本采样（SGNS）的词嵌入方法本质上可以看作词的点互信息（PMI）矩阵的分解形式，若将语料库中词频记为#(·)，则目标词与对应的上下文词向量的点积 $\vec{w}_i \cdot \vec{c}_j$ 与两个词的点互信息存在如下关系：

$$M_{ij}^{SGNS}=\vec{w}_i\cdot\vec{c}_j=\log\left(\frac{\#(w_i,c_j)\cdot|D|}{\#(w_i)\cdot\#(C_j)}\right)-\log k$$
$$=PMI(w_i,c_j)-\log k \tag{1}$$

在同一时期，Pennington[8]也提出类似的观点，认为词的共现计数与上述矩阵直接相关，并由此提出 Glove 词嵌入方法，在特定任务中的性能超越了 word2vec。使得基于矩阵分解的词嵌入方法成为主流。Shi[13]借鉴 Levy 等人的理论，提出矩阵协同分解方法用于跨语言词嵌入方法，并通过融入词翻译概率知识，在英语-德语跨语言文档分类中获得优秀表现。

3 模型描述

3.1 点关联测度与词嵌入的矩阵分解

文献[8][9]分别用不同的目标函数验证了矩阵分解方法学习得到的词向量可以有效表示语言中的词义相似度，两者虽在目标函数上存在差异，但最终词嵌入性能接近。本文认为两者是同质的。为此，我们定义词的点关联测度为：一个自然语言词汇与其关联词汇之间关联程度的度量，则可导出基于矩阵分解词嵌入的统一模型。

对于某种语言的语料库 L，其中出现的所有词构成目标词集合记为 $w_1^m=\{w_1,\cdots w_i,\cdots w_m\}$，与 w_1^m 中任意词存在特定关系的关联词集合记为 $c_1^n=\{c_1,\cdots c_j,\cdots c_n\}$，目标词和关联词的词向量矩阵分别记为 $W_{d\times m}$ 和 $C_{d\times n}$，d 为向量维度，m、n 为词典规模。矩阵 $R_{m\times n}$ 表示 w_1^m 和 c_1^n 中词语之间的点关联测度矩阵，且认为该矩阵可以近似分解为 $W_{d\times m}$ 和 $C_{d\times n}$ 的乘积，即满足:

$$R\approx W^T\cdot C \tag{2}$$

基于矩阵分解的词嵌入就是将式(2)作为依据训练 $W_{d\times m}$ 和 C，目标函数的核心部分通常为点关联测度的差值：

$$J=\frac{1}{2}\left\|w_i^T\cdot c_j-r_{ij}\right\|^2 \tag{3}$$

其中，r_{ij} 表示目标词 w_i 和其关联词 c_j 的点关联测度。该目标函数进一步由随机梯度下降算法迭代训练，最终得到训练结果。

式(3)可理解为矩阵分解词嵌入学习的统一模型，文献[8][9]中的目标函数均为其特定形式。如定义 r_{ij} 为词的点互信息(PMI)，则目标函数等价于式(1)。如果定义 r_{ij} 为 w_i 和 c_j 的共现计数，则近似于文献[8]提出的目标函数。根据该模型，基于矩阵分解方法学习词嵌入必须解决 3 个核心问题：

(1)如何根据语料数据，确定目标词和对应的关联词；

(2)选取何种点关联测度作为训练参数的依据；

(3)如何设计合理的目标函数。

下面将针对以上 3 个问题讨论跨语言词嵌入问题。

3.2 跨语言关联词的确定

在计算跨语言词嵌入时，不仅要考虑本语言的关联词，还要考虑跨语言的关联词，通常这种关联词可以从双语并行语料中获取。不妨将中英对齐语料记为 $D=\{S,T,A\}$，S 为中文语料，T 为英文语料，A 为两者之间的词对齐信息，中英文词典记为 V^S、V^T，D 中的一个句子对记为 (s^i,t^i,a^i)。我们直接以 V^S、V^T 作为跨语言词嵌入的目标词集合，将目标词对应的

关联词集合记为$\tilde{V}^S$、$\tilde{V}^T$。

下面以中文词为例描述关联词确定方法。可以进一步分为两种情况：

(1)使用词对齐知识。此时，对于句对(s^i,t^i,a^i)中出现的中文词$v_k^S \in V^S$，$v_k^S \in s^i$，其关联词包括两部分：v_k^S在s^i中的上下文词，以及v_k^S在t^i中对齐词$a^i(v_k^S)$的上下文词。如图 1 中例子，中文目标词“照顾”与英文词“caring”对齐，令关联词的窗口长度为 3，则目标词在中、英文中的关联词为图中所有箭头指向的词的集合。

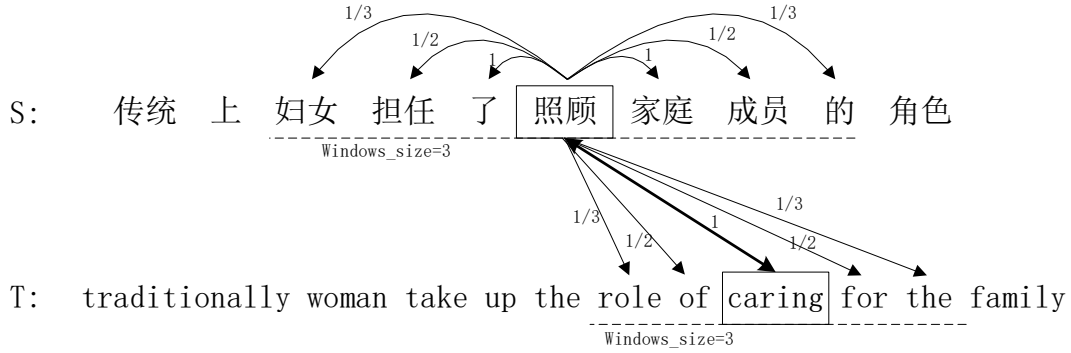


图 1　词对齐情况下关联词的确定示例

(2)不考虑词对齐知识。此时，v_k^S的关联词也包括两部分：在s^i中的上下文词，以及所有在t^i中出现的词，认为v_k^S与t^i中所有词具有平均概率的对齐。

3.3 跨语言关联测度计算

本文分别以跨语言共现计数和跨语言点互信息作为点关联测度，给出其在跨语言词嵌入问题中的计算方法。以中文目标词为例，分两种情况讨论跨语言共现计数的计算方法，目标词为英语词时的计算与之对称：

(1)使用词对齐信息，对于语料库中的句子对$(s^i,t^i,a^i)\in D$，目标词$v_k^S \in s^i$，$\tilde{v}_l^S$是其上下文词，在英文中的对齐词记为$\tilde{v}_h^T = a^i(v_k^S)$，$\tilde{v}_j^T$是$\tilde{v}_h^T$的上下文词。则$v_k^S$与同语言关联词$\tilde{v}_l^S$、跨语言关联词$\tilde{v}_j^T$的加权共现计数分别为：

$$CO(v_k^S,\tilde{v}_l^S)=\frac{1}{|l-k|} \tag{4}$$

$$CO(v_k^S,\tilde{v}_j^T)=\frac{1}{|j-k|}+1 \tag{5}$$

显然，关联词共现计数是根据距离的加权的，距离目标词越远则权重越低。图 1 给出关联词的权重计算示例，其中跨语言的直接对齐词距离从 1 开始计算。

(2)不使用词对齐信息，$CO(v_k^S,\tilde{v}_l^S)$与式(X)中一致。v_k^S与对齐的英文句子中任意词$\tilde{v}_j^T$的共现计数由下式计算：

$$CO(v_k^S,\tilde{v}_j^T)=\#(v_k^S)\cdot\#(\tilde{v}_j^T) \tag{6}$$

其中$\#(v_k^S)$、$\#(\tilde{v}_j^T)$分别是v_k^S、$\tilde{v}_j^T$在s^i、t^i中出现的次数。最终的共现计数由所有句子中共现计数累加得到，不再另行列出。

跨语言点互信息的计算依赖于共现计数，对于两种语言均采用相同的形式，有：

$$CP(v_k^S,\tilde{v}_j^{S,T})=\frac{|D|\cdot CO(v_k^S,\tilde{v}_j^{S,T})}{CO(v_k^S,\bullet)\cdot CO(\bullet,\tilde{v}_j^{S,T})} \tag{7}$$

其中 $\tilde{v}_j^{S,T}$ 是 v_k^S 的关联词，$|D|$ 是双语中出现的所有计数之合，有：

$$|D| = \sum_{S,T} CO(v_p^{S,T}, \tilde{v}_q^{S,T}) \tag{8}$$

在计算过程中，为保证两种语言计算得到的共现概率处于同一个概率空间，计算 $|D|$ 时不区分语言。

3.4 目标函数设计

跨语言词嵌入的两种语言的词向量处于同一个 d 维向量空间中，其中目标词集合为 $V = \{V^S, V^T\}$，关联词集合为 $\tilde{V} = \{\tilde{V}^S, \tilde{V}^T\}$，则采用跨语言计数作为关联测度时的目标函数为：

$$J_{CO}(D,V,\tilde{V}) = \frac{1}{2}\sum_{\substack{v_i \in V \\ \tilde{v}_j \in \tilde{V}}} f(CO_D(v_i,\tilde{v}_j))(v_i^T\tilde{v}_j + b_i + \tilde{b}_j - CO_D(v_i,\tilde{v}_j))^2 \tag{9}$$

类似地，采用跨语言点互信息作为关联测度时的目标函数为：

$$J_{CP}(D,V,\tilde{V}) = \frac{1}{2}\sum_{\substack{v_i \in V \\ \tilde{v}_j \in \tilde{V}}} f(CO_D(v_i,\tilde{v}_j))(v_i^T\tilde{v}_j + b_i + \tilde{b}_j - CP_D(v_i,\tilde{v}_j))^2 \tag{10}$$

其中，b_i 和 $\tilde{b}_j$ 分别为目标词和关联词对应的词向量偏置，$f(x)$ 为低频共现的惩罚权重函数，当目标词与关联词共现计数过低时，损失函数置信度降低，故降低其权重。本文使用 Glove[8] 中给出的权重函数，如：

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha}, & if\ x < x_{\max} \\ 1, & if\ x \geq x_{\max} \end{cases} \tag{11}$$

在训练过程中，除了使用对齐语料库外，还可以将非对齐的单语语料库作为额外的训练数据加入训练过程，以强化词嵌入的表示能力。将参与训练的单语语料记为 S'、T'，则(8)、(9)两种损失函数可扩展为如下形式：

$$J_{CO-Total} = J_{CO}(D,V,\tilde{V}) + J_{CO}(S',V,\tilde{V}) + J_{CO}(T',V,\tilde{V}) \tag{12}$$

$$J_{CP-Total} = J_{CP}(D,V,\tilde{V}) + J_{CP}(S',V,\tilde{V}) + J_{CP}(T',V,\tilde{V}) \tag{13}$$

4 跨语言词嵌入的应用和评价

4.1 跨语言文本分类

假设两种语言 W，T 的跨语言词向量矩阵记为 $W_{u\times d}$、$T_{v\times d}$，$D_m^T = \{(c_1, d_1^T), \cdots, (c_m, d_m^T)\}$ 是由 m 个 T 语言文档组成的集合，其中 c_i 是文档 d_i^T 对应的类型标签。$D_n^W = \{d_1^W, \cdots, d_n^W\}$ 是 n 个 W 语言文档组成的集合，没有对应的类型标签。任意文档的表示向量可由文档出现词的向量经 tfidf 加权累加得到。则 CDC 问题可描述为：以 D_m^T 的表示向量作为训练集训练感知分类器，对 D_n^W 中的文档进行分类，用该分类器的分类正确率来评价跨语言词嵌入的性能。

CDC 问题通常使用 NIST 发布的 Reuters RCV1/RCV2 文档集[1]作为 CDC 数据源。其中 RCV1 为英文文档集，RCV2 为多语言文档集，共包含 4 类文档，每个文档均对应于一个或多个类型标记。与之类似，本文从 RCV1 和 RVC2 中选取具有单一类型的英文、中文文档。考虑到

[1] http://trec.nist.gov/data/reuters/reuters.html

RCV2 中提供的中文文档数量远小于 RCV1 的英文文档数量，我们对英文文档数量做了随机选取，以保证两种语言文档数量、词汇数量基本协调。最终抽取得到中文文档共 24330 个，词典规模为 10.56 万词；英文文档 33286 个，词典规模为 13.65 万词。详细的数据使用情况将在实验部分介绍。

4.2 跨语言词义相似度计算

词义相似度是评价词嵌入的重要方法，英文词嵌入评价常以 WordSimilarity-353（WS353）[16]作为测试集。WS353 数据中包含 353 对英文单词，由 10 名以上标注者对每对单词的相似度进行 1～10 分打分，取平均值作为最终相似度。该相似度与待测试词向量计算得到的相似度取 Spearman 相关系数，作为评价词嵌入的指标。

本文将该方法应用于跨语言词嵌入的评价。由于目前没有公开的中英词汇相似度数据集，我们首先对 WS353 进行翻译，将其中所有单词翻译为中文，并沿用原始相似度打分。例如，对于实例 (e_1,e_2,s)，s 为人工标注的相似度数值。经人工翻译后扩展为 (e_1,c_1,e_2,c_2,s)，则该实例的跨语言词义相似度计算公式记为：

$$Sim(e_1,c_1,e_2,c_2)=\frac{Sim(e_1,c_2)+Sim(e_2,c_1)}{2} \quad (14)$$

在翻译过程中，尽量使用短词，以减少翻译倾向性对翻译结果的影响。仍以 Spearman 相关系数作为最终的评价指标。

5. 实验和分析

5.1 训练跨语言词嵌入

本文中用于训练跨语言词嵌入的数据包括两部分：（1）NIST2008 机器翻译评测提供的中英双语对齐语料；（2）从 RCV1/RCV2 抽取文档集合，其中的双语数据是非对齐的。在预处理阶段，我们将英文语料全部小写化，采用 LTP[2]工具对中文进行分词，去掉双语中所有常规标点和特殊符号，去掉过长和过短的句子。预处理后，中英对齐语料共计 425 万句对，包含中文词 61.39M，英文词 72.22M，使用 SymGiza++[17]工具学习其中对齐知识。作为对比，我们采用类似 Trans-gram[14]提出的方法训练跨语言的词向量表示：利用 SymGiza++生成的对齐语料，以 p=0.5 的概率进行随机混合，然后以混合后的语料作为训练数据，使用 Glove 工具训练词嵌入。

实验主要考察不同的知识和目标函数对跨语言词向量性能的影响，分为 3 个维度：（1）采用何种点关联特征作为学习目标；（2）在训练中是否使用词对齐知识；（3）在训练中是否使用非对齐单语语料。训练过程中的其他的参数设置参考了文献[1][13]，包括：低频权重调整参数 x_max_monoligual=30.0,x_max_bilingual=100.0，α=0.75；低频词截止参数 min_count=5；学习率 η=0.05；共现窗口长度 window_size=10；训练迭代次数为 50。实验中，每种词嵌入方法均训练 $\dim=\{20,40,80,160\}$ 四种维度，以观察维度的变化对性能的影响。

5.2 跨语言文档分类实验

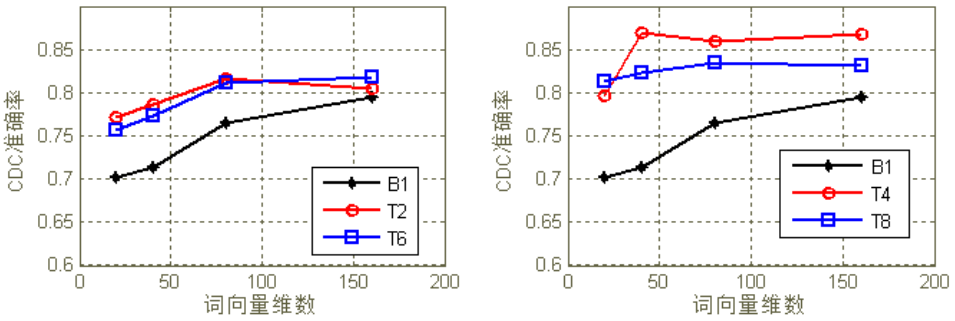
如前文所述， RCV1/RCV2 文档集作为非对齐语料库参与词嵌入训练。在 CDC 实验中，首先利用该语料库计算中英文单词的 IDF 权重，然后根据文档集中 4 种类型文档的原始分布比例随机抽取 10000 个英文文档作为训练集，抽取 5000、1000 个中文文档作为测试集、参数调试开发集。我们在不同参数下训练跨语言词嵌入，使用感知器算法学习跨语言文档分类知识，测试其性能。设置感知器算法迭代次数为 10 次。

[2] http://www.ltp-cloud.com/

实验设计两组对比结果，其中 B0 是测试集中文档数量最多的类别的比例，B1 是以随机混词为训练数据得到的词嵌入的性能测试。实验文档的分类正确率作为评价标准，最后的测试结果汇总于表 1，其中用于 CDC 实验的所有向量的维度为 d=40。

在 CDC 问题中，采用简单的随机混词方法能够得到较好的性能。本文训练得到的词向量在 CDC 问题上的性能均超过 B1。根据表 1 可以得到如下结论：首先，使用非对齐的单语语料参与训练，会使得分类正确率有较大提升，最高有 T4-T3=9.06%。这说明在词嵌入训练过程中，跨语言的知识可以与领域知识分开表达，在有一定规模对齐语料的前提下，使用非对齐的领域语料可以提升特定任务下跨语言词嵌入的性能。其次，使用词对齐信息时的性能普遍高于未使用对齐信息的情况，如 T4-T2=8.35%，T7-T5=5.92%等。这是因为在 CDC 问题中，文档由高权重关键词的向量加权表示，高权重词的对齐特征更加重要。

此外，表 1 中基于跨语言共现计数（CO）和基于跨语言点互信息（CP）两种关联测度下得到的词嵌入性能差别不大，但 T4、T8 的性能有明显提高。为解释该现象，我们分别在 20,40,80,160 四种维度重复训练词嵌入，并进行测试得到对比结果如图 2。



(a) (b)

图 2 不同方法、不同维数词嵌入在 CDC 上的性能对比

图 2(a)是 CO 方法与 B1 的对比结果，在 d=40、80 情况下 T2 略高于 T6；图 2(b)是 CP 方法与 B1 的对比结果，在 d=40、80 情况下 T4 高于 T8。两种方法均在 d=40、80 时取得最高准确率。说明对于 CDC 问题，基于跨语言共现计数的方法更有效，原因是互信息倾向于选择低频词，使得文档表示时低频词权重过高，反而影响最终性能。

表 1 跨语言词向量实验结果

分组	ID	实验参数说明	CDC 正确率(%) d=40	CLSim 相关度 d=80
Baseline	B0	最多文档数分类占比	58.90	-
	B1	随机混词词嵌入	71.34	0.4101
CO（以跨语言共现计数作为学习目标）	T1	无对齐信息，仅对齐语料	74.95	0.4367
	T2	无对齐信息，对齐语料+非对齐语料	78.69	0.4654
	T3	有对齐信息，仅对齐语料	77.98	0.4002
	T4	有对齐信息，对齐语料+非对齐语料	**87.04**	0.4010
CP（以跨语言 PPMI 作为学习目标）	T5	无对齐信息，仅对齐语料	74.28	0.4713
	T6	无对齐信息，对齐语料+非对齐语料	77.43	**0.4801**
	T7	有对齐信息，仅对齐语料	80.20	0.4483
	T8	有对齐信息，对齐语料+非对齐语料	82.50	0.4603

5.3 跨语言词义相似度实验

表 1 的最后一列给出跨语言词义相似度（CLS）的实验结果。由于对比实验 B1 在 d=80 时取得最好性能，因此该实验以 d=80 时的实验结果做比较。

实验中，跨语言互信息学习得到的词嵌入明显好于共现计数，最高有 T8-T4=0.593，该结果与 CDC 的实验结果恰好相反，说明点互信息更适合于表示词与词之间的相似性，而不是文档级别的相似性。

其次，在词嵌入过程中不使用词对齐信息，在测试中取得更好地性能，如 T1-T3=0.365，T5-T7=0.230。预训练得到的词对齐信息虽然更加精确，但相对于无词对齐时的平均分布，仍存在信息损失。也说明在词相似度计算方面，使用更多的关联词能够得到更好地效果。最后，非对齐的语料对词相似度的计算仍然有贡献，说明单语的语料虽然不包含跨语言知识，但作为补充数据仍然有助于词向量性能的提升。

同样，为了考察不同维度词嵌入在 CLS 问题中的变化趋势，我们以 B1、T2、T6，T4、T8 为例进行测试，结果如图 3 所示。各组实验都在 d=80 时取得最好性能。图 3(a)中，由于训练 B1 使用的混词语料也使用了词对齐信息，且同样采用共现计数训练，因此 T4 与 B1 性能非常接近，三者趋势基本一致，无词对齐情况下性能更好。图 3(b)中，基于互信息的词嵌入整体趋势与 B1 类似，但性能有较大提高，说明该方法的有效性。

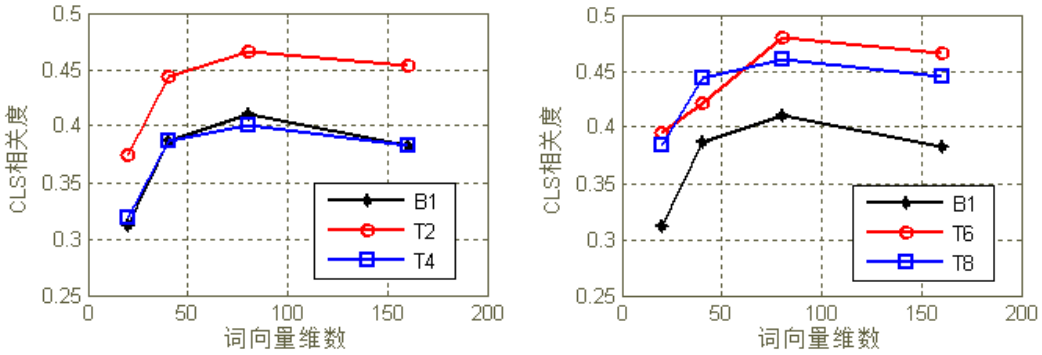


图 3 不同方法 CLS 性能对比

表 2 给出跨语言词嵌入实验 T6 的相似词计算示例，目标词分为中文和英文两组，选取相似度最高的 5 个跨语言词作为示例。可见，目标词与其 Top1 相似词非常接近直译结果。对比而言，根据中文词汇计算英文相似度的结果相对较好。

表 2 跨语言词向量 CLS 性能

Top5	经济	足球	总统	非常	逐渐	欺骗	调查
1	economic	football	presidential	very	gradually	deceive	survey
2	economy	soccer	president	extremely	rapidly	deception	investigation
3	financial	tennis	bush	quite	steadily	mislead	surveys
4	growth	basketball	clinton	indeed	increasingly	cheat	conducted
5	development	badminton	putin	most	slowly	cheated	findings
Top5	economy	football	president	very	gradually	cheat	Investigate
1	经济	足球	主席	非常	逐渐	欺骗	调查
2	复苏	篮球	总统	很	逐步	骗	查明
3	发展	足球队	先生	十分	渐渐	出卖	投诉
4	增长	棒球	女士	相当	慢慢	杀	制止
5	金融	球迷	议员	极	日渐	说谎	追究

5.4 单语词义相似度实验

跨语词嵌入不仅能够表示两种语言之间词汇关联特征，而且在各自语言中也应该具有词嵌入的基本特征。本文使用英文 Sim353 测试集，对跨语言词嵌入得到的英文词向量进行测试。首先用 Glove 工具，仅使用双语对齐语料中的英文数据单独训练 d=80 维的英文词嵌入，并对 Sim353 数据集，计算整体 Spearman 相关度，作为参考记为 B2。使用由同样数据训练的词向量（T1、T3、T5、T7）进行测试，结果见表 3。

表 3 单语词嵌入相似度实验

分组	编号	Sim353 相关度	
Baseline	B2	0.4562	
CO 为目标	T1	0.4620	↑
	T3	0.4478	↓
CP 为目标	T5	0.4644	↑
	T7	0.4645	↑

可见，在相同的训练数据条件下，除了 T3 性能略有下降，其余几组实验均超过了 Glove 训练得到的词嵌入，证明本文所述方法的有效性。

6 结论

近几年，词嵌入在自然语言处理中扮演了日益重要的角色。以特定关联测度为学习目标，借鉴矩阵分解形式设计机器学习系统，是获取词嵌入的主要方法。本文将该方法扩展到跨语言词嵌入训练问题中，以对齐语料为主要知识源，分别探讨跨语言共现计数和跨语言点互信息作为词的关联测度情况下跨语言词嵌入的训练方法。提出跨语言关联词确定方法和统一形式的词关联测度的计算方法，建立目标函数实现词向量的学习。

本文采用跨语言文档分类和跨语言语义相似度计算作为评价词嵌入的主要依据，从多个层面测试影响跨语言词嵌入性能的因素。通过实验验证跨语言共现计数作为关联测度适合于解决 CDC 问题，而跨语言点互信息则适合于解决 CLS 问题。训练得到的跨语言词嵌入能够有效表示中英文词汇之间的语义联系，同时其词向量在单一语言中能够保持性能不降低。因此该方法可以作为跨语言信息处理的预处理表示方法，广泛应用于各类应用中。

最后，本文提出的基于点关联测度的词嵌入方法，研究新的点关联测度，针对特定任务、特定领域使用特定的点关联测度，可以提高词嵌入训练的灵活性和性能，将是本工作未来研究的重点。

参考文献

[1] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words[C]// In Proceedings of COLING 2012: Technical Papers. Mumbai, 2012:1459-1474.

[2] Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations[C]// In Proceedings of NIPS2014. Montreal, 2014:1853－1861.

[3] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation[C]// In Proceedings of EACL2014. Gothenburg, 2014:462－471.

[4] Ivan Vulic and Marie-Francine Moens. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction[C]// In Proceedings of ACL2015(Short papers). Beijing, 2015:719-725.

[5] Stephan Gouws, Anders Sogaard. Simple task-specific bilingual word embeddings[C]// In Proceedings of NAACL2015. Denver, 2015:1386-1390.
[6] Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation[C]// In Proceedings of EMNLP2013. Seattle, Washington, 2013:1393–1398.
[7] Jiang Guo, Wanxiang Che, David Yarowsky, HaifengWang, and Ting Liu. Crosslingual dependency parsing based on distributed representations[C]// In Proceedings of ACL2015.Beijing, 2015:719-725.
[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation[C]// In Proceedings of EMNLP2014. Doha, 2014:1532–1543.
[9] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization[J]. Advances in neural information processing systems. 2014(3):2177-2185.
[10] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics[C]// Eprint Arxiv, arXiv:1312.6173v4.
[11] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In Proceedings of ICML2015. Lille, 2015:748–756.
[12] Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. Leveraging monolingual data for crosslingual compositional word representations[C]// In Proceedings of ICLR2015. San Diego, 2015.
[13] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingualword embeddings via matrix co-factorization[C]// In Proceedings of ACL2015(Short papers). Beijing, 2015:567-572.
[14] Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, Fast Cross-lingual Word-embeddings[C]// In Proceedings of EMNLP2015. Lisbon, 2015:1109-1113.
[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality[C]// In Proceedings of NIPS2013. South Lake Tahoe, 2013:3111–3119.
[16] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing Search in Context: The Concept Revisited[J]. ACM Transactions on Information Systems, 2002,20(1):116-131.
[17] Marcin Junczys-Dowmunt and Arkadiusz Szat. Symgiza++: symmetrized word alignment models for statistical machine translation[J]. In Security and Intelligent Information Systems, 2011(7053):379-390.

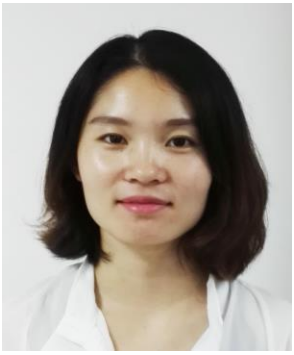
通讯作者联系方式：于东　北京市海淀区学院路 15 号北京语言大学信息科学学院 邮编 100083

电话：13811102410 电子邮箱 yudong_blcu@126.com

作者简介：

于东（1982-），男，博士，副教授，主要研究方向为自然语言处理；

赵艳（1994-），女，硕士研究生，主要研究方向为语言信息处理；



韦林煊（1995-），男，本科生，主要研究方向为语言信息处理；

荀恩东（1967-），男，博士，教授，主要研究方向为语言信息处理、教育技术。