基于跨语言语料的汉泰词分布表示

张金鹏[1,2]，周兰江[1,2]，线岩团[1,2]，余正涛[1,2]，何思兰[3]

(1.昆明理工大学 信息工程与自动化学院，云南 昆明 650500；

2.昆明理工大学 智能信息处理重点实验室，云南 昆明 650500：

3.昆明理工大学 理学院，云南 昆明 650500)

摘要：词汇的表示问题是自然语言处理的基础研究内容。目前单语词汇分布表示已经在一些自然语言处理问题上取得很好的应用效果。然而在跨语言词汇的分布表示上国内外研究很少，本文针对这个问题，利用两种语言名词、动词分布的相似性，通过弱监督学习扩展等方式在中文语料中嵌入泰语的互译词、同类词、上义词等，学习出泰语词在汉泰跨语言环境下的分布。实验基于学习到的跨语言词汇分布表示应用于双语文本相似度计算和汉泰混合语料集文本分类，均取得较好效果。

关键词：弱监督学习扩展，跨语言语料，跨语言词汇分布表示，神经概率语言模型

中图分类号：TP391 **文献标识码：**A

cross-lingual corpus-based word distributed representations of Chinese and Thai

ZHANG Jinpeng[1,2],ZHOU Lanjiang[1,2], XIAN Yantuan[1,2], YU Zhengtao[1,2],HE Silan[3]

(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500，China;

2. The Key Laboratory of Intelligent Information Processing,Kunming University of Science and Technology, Kunming, Yunnan 650500，China;

3. The School of Science, Kunming University of Science and Technology, Kunming, Yunnan 650500，China)

Abstract: Word representations is the basic research content of natural language processing, At present, monolingual word distributed representations has shown satisfactory applications effect in some NLP research.While on the cross-lingual word distributed representations , there is little research both at domestic and abroad. Aiming at this problem, we embed mutual translated word,synonyms,superordinates into chinese corpus by utilizing noun and verb distribution similarity of these two languages,weakly supervised learning extension and other methods, learned the thai word distribution in cross-lingual environment of Chinese and Thai. We applied the cross-lingual word distributed representations that learned before to compute bilingual text similarity and the classification of Chinese and Thai mixed text corpus,experiment shows that these two work both have a good effect.

Keyword:weakly supervised learning extended; cross-lingual corpus;cross-lingual word distributed representations;Neural Probabilistic Language Model

1. 引言

词汇表征问题是自然语言处理的重要内容，是信息检索、数据挖掘、知识图谱构建等研究方向的重要技术支持。基于统计机器学习的词汇表征方法的目标是从自然语言文本中学习出词序列的概率表示函数，其面临的一个困难在于词向量的维度灾难与数据稀疏问题[1]，在训练的过程中每一个词序列与其它训练语料中的词序列在离散空间表示时有很大的不同。

收稿日期：2015-6-15 **定稿日期：**2015-8-10

基金项目：国家自然科学基金项目面向互联网的泰语-汉语双语语料获取及对齐方法研究（61363044）

在单语词汇的空间表示过程中，一个传统但有效的方法是 n 元语法模型，它通过学习目标词汇一个短的窗口信息来预测目标词汇出现的概率。它的缺点在于不能反映窗口以外的词对序列生成概率的影响及相似词序列的分布概率的相似性[2]。Bengio 等人在 2001 年提出的神经概率语言模型在单语环境中较好地解决了这个问题。神经概率语言模型通过从自然语言文本中获取句法语义信息学习出词语的分布表示特征，对相似的词序列有相似词分布，Collobert R[3]等验证了词分布能很好地应用于词性标注、命名实体识别、语义角色标注等自然语言问题。虽然单语词汇分布表示上取得了不错的效果，但在跨语言自然语言处理领域的国内外研究稀少，目前主要有两种方法：第一种是迁移学习[4-7]，该方法将标记学习信息从一种语言迁移到另一种语言，使得资源较少的语言获得较好的处理效果。Zeman 等在跨语言句法树库建设上验证了该方法[4]，但该方法有较大的局限性，其效果直接依赖于知识转移的过程，不同的跨语言自然语言处理任务有不同的迁移方法；第二种方法将两种语言转化为其中一种语言或第三方语言上，用一种语言表达跨语言信息[8-10]。Steinberger 等在跨语言文本相似度计算上应用了该方法[8]，但这些方法无一例外依赖已有的双语翻译概念词典（如 WordNet）的质量或统计语料共现信息来计算跨语言词之间的相似度，需要解决译词歧义问题，过程复杂，效果有限。

以上方法在处理跨语言自然处理问题上都取得了一定的效果，但都存在可移植性不强，算法过程复杂，准确性存在提升空间的缺点。目前主流的文本层面分析方法只考察名词、动词的分布特征，借鉴这一思想，本文针对以上存在的问题分析汉语、泰语名词、动词的分布相似性，将泰语名词、动词看做汉语名词、动词，将泰语词嵌入到汉语语料中，生成汉泰跨语言词汇序列语料，通过神经概率语言模型学习泰语名词、动词在跨语言空间中的分布。通过这种方式将在跨语言语料中学习得到的汉泰跨语言词向量分布表示，直接应用到泰语文本，解决泰语学习语料资源缺少和跨语言文本分析问题。本文基于汉语、泰语跨语言文本分类和文本相似度实验，验证了汉语、泰语跨语言词汇分布表示的效果。

本文第 2 节介绍了神经概率语言模型，第 3 节介绍了汉语泰语跨语言语料生成方法，第 4 节对本文的方法进行了测试与评价。

2. 神经概率语言模型

神经概率语言模型[2]由 Bengio 等人于 2003 年第一次提出，基于人工神经网络来学习一种语言的词汇序列的联合概率函数，目前已经在自然语言处理各个领域得到了广泛应用，并取得不错的效果。该模型同时学习每个词的分布和表示词序列的概率函数。模型可以得到泛化是因为一个从未出现的词序列，如果它是由与它相似的词（在其附近的一个代表性的意义上）组成过已经出现的句子的话，那么它获得较高的概率。它有效得解决了词典向量语言空间的维度灾难与数据稀疏问题，同时解决了 n 元语法模型不能解决的分布相似问题，从而比词典向量及 n 元语法模型更好地表示词汇的分布。

神经概率语言模型的描述如下：

通过给定的词序列 $w_1 \ldots\ldots w_t$,其中 $w_t \in V$ ， V 代表目标语言所有的词汇集， V 虽然很大但有限，神经语言模型的目标是要学到一个好的函数来估计词汇的条件概率：

$$f(w_t, w_{t-1} \ldots\ldots w_{t-n+2}, w_{t-n+1}) = p(w_t \mid w_1^{t-1}) = p(w_t \mid w_{t-1}, w_{t-2} \ldots w_1) \qquad (1)$$

其中， w_t 表示词序列的第 t 个词；把相应的子序列写成 $w_i^j = w_i, w_{i+1} \ldots w_{j-1} w_j$; V 表示词表，$|V|$表示词表的大小。通过条件概率的可以获得词序列的联合概率。

公式 1 包括两个过程：

首先构建映射 C 将词汇集V 中的任意元素映射到词的特征向量 C(i) ∈R^d，它代表关联词表中词的分布特征向量。d 代表特征向量的维度。实验中被表示为$|V|*d$ 的自由参数矩阵。

然后构建词的概率函数。我们用词的特征向量 C 表示：根据语料库词的输入上下文特征向量($C(w_{t-n+1})…C(w_{t-1})$)预测在词表中下一个词的概率分布。g 的输出是一个向量，向量的第 i 个元素为估计概率为 $\hat{p}(wt = i \mid w1^{t-1})$。通过如下方式计算：

$$f(i, w_{t-1}.....w_{t-n+2}, w_{t-n+1}) = g(i, C(w_{t-1})...C(w_{t-n+2})C(w_{t-n+1}) \quad (2)$$

f 由以上映射 C 与 g 组合而成，这两个映射都关联一些参数。映射 C 的参数就是特征向量本身，被表示成一个$|V|*d$ 的矩阵 C，C 的第 i 行是词 i 的特征向量。函数 g 可由前馈神经网络或卷积神经网络实现。上式表明函数 f 通过上下文词来预测词表中第 i 个词最终转化为函数 g 通过上下文词的分布特征向量来预测第 i 个词的分布。

3. 汉泰跨语言词分布表示

3.1 汉语与泰语的词序列分布特点

汉语与泰语有较大程度的相似性，它们在语法上有很多共同点。例如针对同一句话：汉语的句法结构为（+定语）主语+（+状语）谓语+（+定语）宾语（+补语）；而泰语的句法结构为主语（+定语）+谓语+宾语（+定语）（+状语或补语），两种句子的主干：主谓宾序列关系完全一致，主要差异体现在泰语的定语、状语必须放在中心词之后，而汉语的定语、状语必须放在中心词之前。从句子的组成来讲，主干反映句子的主要内容，定状补是枝叶成分可有可无，两种语言主干主谓宾成分是完全一致的，主谓宾对应词性中的名词、动词，两者句子主干结构一致。两种语言名词、动词的词序列的分布也应该是有相似性的。

正是由于汉语与泰语在以上句子词序列上的主干相似性决定在同一分布空间下用相同维度向量表征名词、动词的分布成为可能，在自然语言处理中，文本分析只考察名词、动词，解决了名词、动词的跨语言词分布问题也就解决了跨语言文本分析问题。

我们目标旨在忽略中泰两种语言的差异，将泰语名词、动词看做汉语名词、动词，在汉语的语言环境下学习它们的分布，从而使较为成熟的汉语的文本分析方法可以直接应用在泰语文本上。

3.2 平行语料预处理

我们选取从中国广播电台获取并人工校正得到平行句对 10216 对。尽管原始文本包含所有的文本信息，但是目前的自然语言处理技术无法完全处理这些文本信息，因此，需要对文本进行预处理。传统的文本预处理主要是去除停用词，如“的”“地”等。由于本文的方法需要对词的序列分布进行学习所以我们没有去除停用词，但我们将一些与汉泰文本内容无关的符号（“#、*”等）、无意义数字去除，并将一些人名等转化为统一的符号，避免因为人名的变化造成对词序列分布学习的影响，减少噪声干扰。

3.3 平行语料词对齐

我们将以上处理后的平行语料输入 GIZA++[11]中，实现汉泰双语词对齐。GIZA++是包含 IBM1-5 训练模型及隐马尔可夫模型的统计机器学习工具包。GIZA++有几种词对齐启发式算法，我们主要使用交叉启发式算法，通过运行从汉语映射到泰语及从泰语映射到汉语两个方向来获取对齐词对。我们只考虑在两个方向都有的对齐词对。通过词对齐我们可以获取一个词语在平行语料中相应的跨语言翻译词。

例句：今天/0 下午/1 我们/2 要/3 打/4 篮球/5 (1)

ช่วงบ่าย/1 วันนี้/0 พวกเรา/2 จะ/3 เล่น/4 บาสเกตบอล/5 (2)

在上面的平行句对中，相同后缀标号的汉泰词表明它们是对齐互译关系。通过 GIZA++ 我们可以得到泰语词在汉语实例中的对应汉语词。我们在汉语实例中将汉语词我们替换为 พวกเรา，生成实例(3)。

今天 下午 พวกเรา 要 打 篮球 （3）

我们将实例(1)、(3)放入神经概率语言模型学习语料中，得到 พวกเรา 词在汉语语言环境境下应有的分布信息，最终可以使 พวกเรา 与汉语词我们的分布相似，符合语言学的规律，即即两个词相似则它们的分布也应相似。我们把像实例(3)这种在汉语实例适当位置嵌入相应泰语词的实例称为衍生实例。将平行句对中的汉语实例及衍生实例一同作为学习语料进行学习可以得到跨语言词汇的分布。

3.4 泰语词与汉语词相似关系替换

一般认为如果两个词之间是词典互译关系就可以认为它们是对齐的，但在更一般情况中，扮演相似语法语义角色的词会有相似的词序列联合概率分布。例如：昆明与云南是部分整体部分关系，昆明可以代表云南，在一些词序列中，云南的位置可以替换为昆明，如云南四季如春可以替换为昆明四季如春。类似的例子如北京可以代表中国，少林寺代表中国佛教文化。另外语义相反也是一种对齐，热与冷虽然是反义，但在很多上下文中词序列分布很相似，在实例今天天气热与今天天气冷中，有关系：P(热|今天，天气) ≈ P(冷|今天，天气)。无论在汉语还是泰语中整体与部分关系，反义关系在词序列中的分布都有一定的相似性，因此我们认为汉语的词如果与泰语的词存在一定的语义关系可以认为汉语词与泰语词之间也是一种对齐关系，例如：汉语的热与泰语的冷（เย็น）我们可以认为是一种可以语义角色互换换的词对，也认为它们是词对齐，我们称为语义词对齐。由于 不同语言版本之间的 WordNet 的同义词集合的 synset_id 是对应的，汉语词与泰语词的语义词对齐可以通过 通过 synset_id 将汉泰文 WordNet 与英文 WordNet 对应，在英文 WordNet 上查询两种语言词的语义关系 。

虽然通过 WordNet 可以考察汉语词与泰语词之间的语义关系对齐，但我们将语义词对齐泛化为更一般的情况，在自然语言词的序列分布中，只要是相似的语法语义角色就会有相似的词序列分布，即词相似则词在自然语言文本的词序列中的分布也相似。例如有以下实例集：(i)很多游客在丽江品尝丽江粑粑。(ii)几个姑娘在树下享用普洱茶。(iii)一群男人在河边吃酸角。对以上三个实例可以进行如图 1 的转化：

从图 1 的三个实例句的成分可以看出，量词很多、一群、几个在句子集中可以相互替换位置，替换后它们各个句子的词序列联合概率分布仍是相似的。反映在神经概率语言模型中条件概率表达为：

$$P(\text{很多} \mid \text{游客}, \text{在}, \text{丽江}, \text{品尝}, \text{丽江粑粑}) \approx P(\text{几个} \mid \text{姑娘}, \text{在}, \text{树下}, \text{享用}, \text{普洱茶})$$
$$\approx P(\text{一群} \mid \text{男人}, \text{在}, \text{河边}, \text{吃}, \text{酸角})$$

同理主语姑娘、男人、游客，状语丽江、树下、河边，动词吃、品尝、享用，宾语酸角、丽江粑粑、普洱茶都可以相互替换。替换过之后的句子词序列联合概率分布仍与原句子序列联合概率分布相似，即两个句子在神经概率语言模型空间上分布表示向量夹角余弦值接近于

1 或者欧氏距离较小。表达为公式为：

$$V(\text{很多,游客,在,丽江,品尝,丽江粑粑}) \approx V(\text{一群,姑娘,在,河边,享用,酸角})$$

由于在平行句对中，与每句泰语平行对齐的汉语句子是其译句，如果泰语句子中的泰语词汇 thwordi 对应的汉语译句中的汉语词 chwordi，而汉语词 chwordi 与其他汉语词 chwordj 存在上例所说的相似对齐，我们认为泰语词 thwordi 与汉语词 chwordj 相似分布对齐。我们将 thwordi 与 chwordj 的这种对齐方式在本文中定义为原理 1。

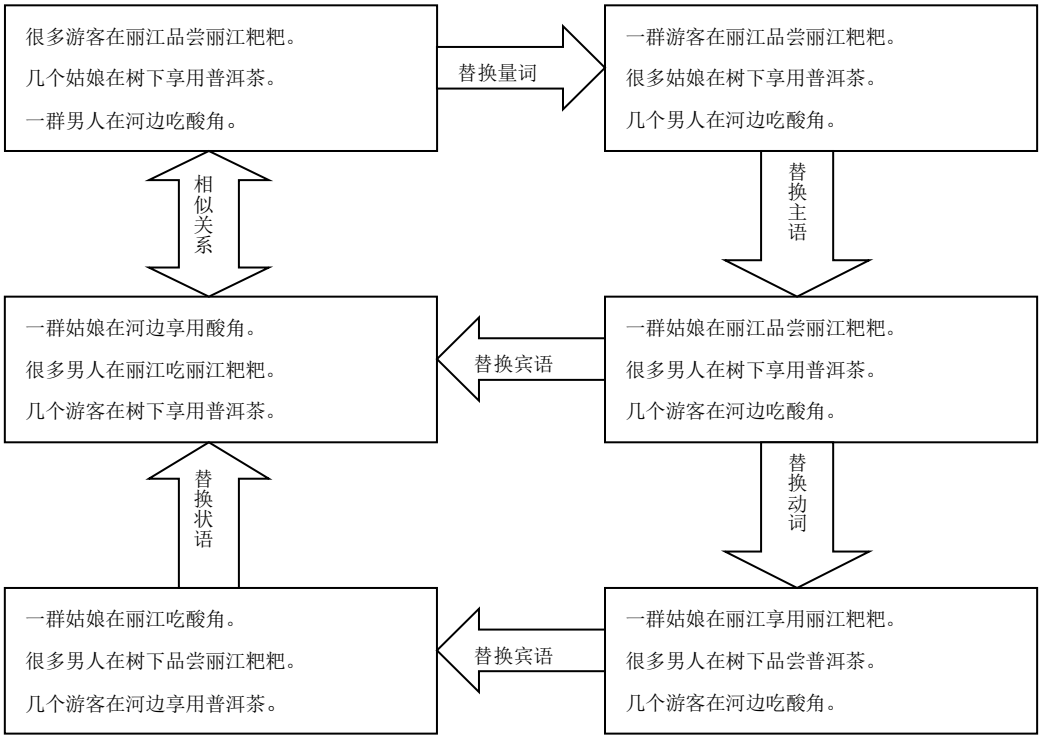

图 1 词序列相似词分布示例

在神经概率语言模型中，相近的词序列会有相近的分布概率表示，由于模型的平滑性，分布的微小改变会影响造成词的预测概率的微小改变，同时训练语料中任何一个实例句子的词的出现不仅会影响这个词的所在句子的联合概率分布也会影响到所有相似（邻近）实例句子的联合概率分布[12]。例如有以下三个实例句：

(1)很多游客在丽江喜欢品尝丽江粑粑。

(2)游客在丽江喜欢品尝丽江粑粑。

(3)很多游客在丽江品尝丽江粑粑。

游客这个词在实例句 1 中的分布概率改变会影响到游客在实例（2）、（3）中的分布概率表示。即游客这个词在实例（1）、（2）、（3）中也是词序列中分布概率表示相似的。实例（1）、（2）、（3）中的任何一个在训练语料中的词序列变化都会影响到其余两个实例最后的概率分布表示。我们认为实例（1）中的游客跟实例（2）、（3）中的游客是相似的。因此，如果泰语句子对应的汉语译句为实例（1），泰语词 thwordi 对应实例（1）中的游客，则 thwordi 与实例（2）、（3）中的游客词也为相似分布对齐。我们在本文中定义这个原理为原理 2。

基于神经概率语言模型的原理 1 与原理 2，我们通过章节 3.3 的工作可以获得泰语句子中的每个泰语词 thowrdi 对应的汉语译句对应的汉语词 chwordi。我们将已经通过神经概率语言模型对平行句对中的汉语句子语料集进行训练得到每个词在汉语语料中分布表示。如果泰语实例句中的泰语词 thwordi 对应的汉语词 chwordi 的分布表示与其他汉语词 chwordj 的

分布表示相近，我们就认为 thwordi 与 chwordj 在神经概率语言模型中相似分布对齐（即它们扮演相似的语法语义角色）。我们把 chwordj 与 thowrdi 的这种相似传播过程称为相似传递。

遍历语料中所有汉语句子，本文通过相似传递统计出每个泰语词 thwordi 与其它汉语词 chwordj(chwordj 不为 thwordi 的互译词)的相似对齐实例。我们通过泰语词 thwordi 与满足相似对齐的其它汉语词 chwordj，在汉语实例中相应位置的 chwordj 替换为泰语词 thwordi 生成新的实例，对语料中的每个泰语名词、动词重复以上过程，直到所有泰语名词、动词都嵌入它们在汉语实例中应有的位置，这个过程中我们不考虑已经在 3.3 中已经计算过的互译对齐词。如果泰语词与汉语词之间的替换衍生实例越多，则说明它们之间的相似程度越高，最后通过神经概率语言模型学习到的词分布越接近。

我们将衍生实例与汉语实例一同作为学习语料进行学习，因为通过衍生实例我们可以学习到泰语名词、动词在汉语语言环境下应有的词分布。通过这个过程使汉语名词、动词与泰语相似的名词、动词有相近的分布，实现汉泰跨语言词汇在同一模型空间下的分布表示。

3.5 大规模汉泰混合语料弱监督学习扩展过程

我们把上节获取的汉语实例与衍生实例混合实例集进行学习可以得到泰语词汇的一定程度汉语表示，但受平行句对所需人工校正参与程度较大，语料规模有限，学习得到的泰语词汇分布还不能完全反映它在汉语语言环境下应有的分布，例如我们可以在混合实例中学习到 พวกเรา 与词我们的相似关系，但 พวกเรา 在汉语中与咱们，咱、俺们等都应是相似的，和他们和他们等代词也有一定相似性。而神经概率语言模型只有在大规模语料集中才能取得最佳效果。因此我们把上文的混合实例集加入到搜狗中文新闻语料集中得到大规模混合语料集。

我们首先在大规模语料集中对泰语名词、动词对应的汉语名词、动词进行替换，这个过程可以通过汉泰版本的 WordNet 同义词相同的 synset_id 来找到泰语词对应的同义汉语词，然后进行替换生成衍生实例，同时保留原语料汉语实例，由于汉语语言的博大精深，一个汉语词可以有多个同义词，我们借助哈工大同义词词林，将泰语词对应替换的汉语词范围扩展到所有的同义词中，如泰语 พวกเรา 除了可以替换我们外，还可以替换我们的同义词如咱、咱咱们、吾侪、吾辈、俺们、我辈、咱俩。通过这个过程我们得到第一次混合语料集。

我们通过同义词替换学习可以得到跨语言学习语料集进行学习，可以取得泰语名词、动词在汉语语言环境下更加准确的序列分布信息。通过这个过程，พวกเรา 与我们、咱们、咱都都取得了相似的分布。

鉴于在神经概率语言模型中，词与词之间如果有相似的上下文语义或者句法构成可以有相近的词分布。基于这种原理，我们认为如果泰语词与语料中的其它汉语词有相似的分布，它们在语料集实例中的位置应该是可以相互替换的，如在实例今天 下午 พวกเรา 要 打 篮球球中的 พวกเรา 可以进一步替换为我、你、你们、他们等同类代词，也可以替换为人名，甚至至可以进一步替换为上义词男人，学生，孩子们等，这样可以使词的分布相似比语义相似更加泛化，也更加符合自然语言文本的分布特性，我们认为这些可以相互替换的代词、人名甚至上义词都符合相容词信息熵值最大的原理。

基于以上分析，我们在第一次跨语言混合语料集中学习得到每个泰语词相似度高于一定阈值的汉语词，并对这些汉语泰语词通过 synset_id 转化为英文，在英语 WordNet 中进行查询，如果汉语词与泰语词属于同类关系或者直接上义词，我们都将泰语词替换汉语词相应的位置，生成新的衍生实例。这个阈值如果选取过高，将很难学习到新的汉语相似词，如果阈值选取过低，则学习得到新的汉语词相似度太低，很多情况下不能替换，我们把阈值设为 0.5。

我们对包含泰语词的语料进行如下过程的弱监督学习扩展：

1. 将泰语词与汉语词的相似度进行比较，如果相似度高于阈值，我们把汉语词放入候选替换词集中；

2. 对泰语词与候选替换词集中的词通过 synset_id 转化为英语，在英文 Wordnet 的 is_a 层级树中查询他们之间的语义关系，如果它们之间是同类词或者直接上义词则可以直接替换，生成新的候选衍生实例；

3. 人工判断候选的衍生实例是否符合知识逻辑，合理则保留该衍生实例，否则就去除该实例。例如实例日本自卫队举行阅兵仪式从语法角度可以替换为 จีน（中国）自卫队举行阅兵仪式。但该新实例不符合知识逻辑；

4. 将筛选出的衍生实例加入语料集中通过神经概率语言模型学习新的汉泰词汇跨语言分布，并跳转到过程 1；

5. 重复过程 1、2、3、4，直到学习不出新的汉语替换词为止。

我们通过以上弱监督学习扩展过程对语料集中的词汇进行学习，我们发现随着语料规模的增大，汉语词与汉语词汇之间的相似度会略有下降，但泰语词汇与汉语词汇之间的相似度明显上升，泰语词汇与汉语词汇之间同义关系的相似度与预期结果较为接近，并接近汉语相近词汇之间的相似度，例如：**การจราจร**（运输的意思）与交通的相似度达到 0.5063 接近运输与交通的相似度 0.5423。

我们将整个汉泰跨语言词汇分布学习过程总结如流程图 2 所示。

我们把汉语神经概率语言模型扩展到汉泰跨语言词汇分布表示上，由于在学习语料中合适的位置嵌入了泰语名词、动词，所以我们的神经概率语言模型经过学习，可以得到汉泰跨语言词汇较为准确的分布表示。

3.6 模型学习

神经概率语言模型中用反向传播算法 [13]学习模型参数。目前针对反向传播算法的参数改进学习算法有很多，我们选用 Zeiler M D[14]等人改进的 ADADELTA 梯度下降算法来最优化模型的参数集。该方法可以动态地适应一阶信息，并对梯度下降有最小的计算开销。训练一次实例就更新一次参数。首先从神经网络的输出层开始，每一层的每个参数的梯度通过后一层的梯度来获得，经过网络的每一层最后到达输入层的词的分布特征向量，不断迭代直至误差符合预期完成整个过程。

4 实验及分析

4.1 文本相似度计算方法

我们首先用神经概率语言对上述跨语言语料进行学习，得到汉泰词汇的跨语言分布表示，基于经验，我们设定每个词的向量维度为 200，神经概率语言模型隐藏层的神经单元个数为 64，允许误差 0.001，训练窗口为 5。在语料集学习的过程中只考虑出现频数大于等于 3 次的汉泰词汇。我们把学习得到的汉泰词汇跨语言分布作为文本相似度计算的基础。

我们通过 tf-idf 算法筛选出每篇文档特征权重占前 5 位的特征词，文本 t 的特征词组为(v_{t1},v_{t2}……v_{t5})，权重为(w_{t1} ,w_{t2},…… w_{t5})，同理文本 k 的特征词组为(v_{k1},v_{k2},……v_{k5})，特征词对应 tf-idf 权重为($w_{k1,}$ $w_{k2.....}$ w_{k5})。两篇文本间的相似度通过文本 t 中的每个特征词与文本 k 中的每个特征词的词向量余弦相似度及各自特征权重的乘积累加求和除以总共相加次数 25。词 v_{k1} 与 v_{t1} 的词向量余弦相似度表示为 $v_{k1}\,\&\,v_{t1}$。文本相似度计算公式为：

$$t\,\&\,k=\frac{\sum_{i=1,j=1}^{5} v_{ti}\,\&\,v_{kj}*w_{ti}*w_{kj}}{25} \tag{3}$$

4.2 实验结果与分析

我们选用维基百科上的汉泰篇章对齐文本作为实验文本集，选取经济、政治、文化、科技、体育五类汉泰平行文本各 100 篇。实验由两部分组成：第一部分：汉泰平行文本相似度计算；第二部分：汉泰混合文本集中的文本随机打乱顺序后判断它们在五大类中的分类。汉泰文本的相似性说明两者之间的同义词的跨语言词分布相似性，只有两篇文本中的同义词在一致的向量空间分布表示上的相似才能使文本相似度高。

维基百科上篇章平行文本都是针对同一词条的描述，但它们在描述上有差异，很多情况下一种语言的描述很详细而另外一种语言描述较简单，我们人工筛选汉泰平行文本描述一致，篇幅相当的文章，经语言学家判定相似程度高于 95%的平行文章。由于我们不追求单

语言环境下的文本相似度效果，只追求在同种计算方式下的双语平行文本相似性，因此采用上节描述的文本相似度计算方法计算相似性。实验结果如下表 1 所示：

表 1 跨语言文本的相似程度

领域	跨语言词汇分布表示平均相似度	实际相似度
经济	81.19%	95% 以上
政治	75.6%	95% 以上
文化	73.02%	95% 以上
科技	69.84%	95% 以上
体育	78.77%	95% 以上

文本相似度实验表明通过跨语言词汇分布表示来表征汉泰文本相似度方面有一定的效果，针对平行文本均取得了 69.84%以上的相似度。

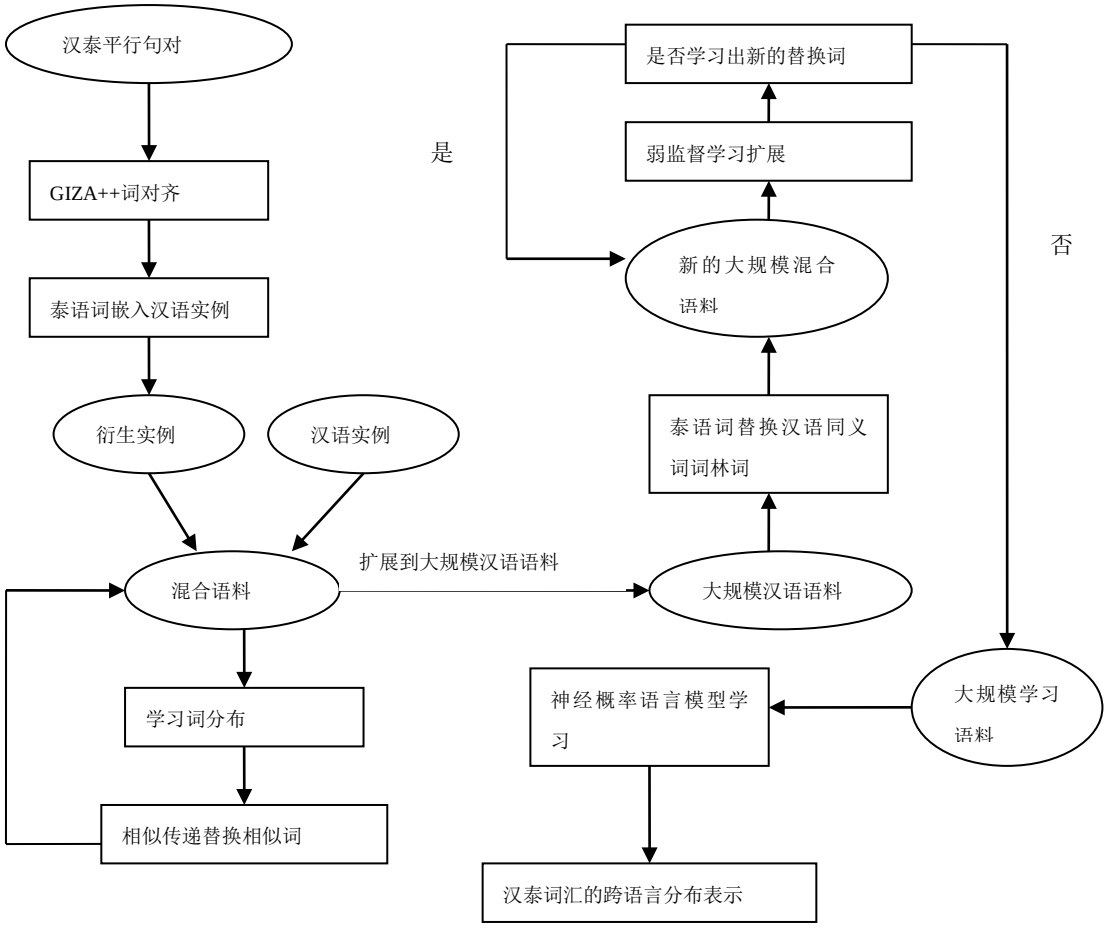

图 2 汉泰跨语言词汇分布学习流程图

汉泰文本混合文本集的文本分类准确性说明汉泰词在跨语言模型中的词汇分布表征准确性，如果词汇的跨语言词汇分布表示不准确会导致文本分类准确率下降。我们的目的是检验跨语言词汇表示的准确性，故我们采用 KNN 文本分类算法，它是较理想的文本分类算法。待分类文本与训练文本相似度计算时采用上节的文本相似度计算方法。我们选取的汉泰文本都是单种分类标记的文本，不考虑多分类标记文本，并将我们的方法同跨语言文本分类效果较好的模型翻译 [15]（通过期望最大算法把源语言分类标记文本翻译为目标语言分类标记文本学习分类知识后分类）、结合半监督适应的模型翻译 [15]（模型翻译同时结合半监督学习更新目标语言的分类特征词分布）及机器翻译（两种方法：1.源语言分类标记文本翻译为目标语言，目标语言待分类文本学习分类知识后分类；2.目标语言翻译为源语言学习分类知识后分类）的方法作对比。结果如表 2 所示：

表 2 跨语言文本分类准确性

领域 语言/方法	经济	政治	文化	科技	体育
跨语言词汇分布表示	79.47%	73.98%	69.62%	61.34%	78.62%
机器翻译（汉-> 泰	64.39%	63.18%	57.6%	52.27%	59.24%
机器翻译（泰-> 汉	63.27%	60.18%	54.6%	48.09%	58.33%
期望最大翻译	81.29%	76.4%	72.03%	68.66%	82.54%
期望最大翻译+半监督适应	87.7%	78.65%	80.18%	84.83%	89.2%

实验结果表明：相同语料规模情况下，跨语言词分布在跨语言文本分类方面较两种机器翻译方式效果较好，略差于基于期望最大算法翻译分类方式，距期望最大算法翻译+半监督适应方式有一定的差距。原因在于跨语言词汇分布可以反映跨语言词汇相似程度，比机器翻译的翻译结果提高了准确性，但分类效果略差于期望最大翻译，因为期望最大算法考虑了在类别信息下源语言词翻译为目标语言词的最大翻译概率，相比跨语言词汇相似度是所有类别下的平均相似度，准确性更高，而结合半监督适应后可以更新目标语言文本分类的分布特征词，效果更好。实验说明汉泰跨语言词汇分布表示上的准确性，即词汇意义的表达准确性。本文的方法在跨语言文本分类方面效果不是最佳但过程简单，基于跨语言词分布将源语言的分类知识直接迁移到目标语言上，有一定效果的同时速度最快。

5 结论及展望

本文为解决汉泰词汇的跨语言分布表示问题，忽略两种语言的差异，将泰语名词、动词嵌入到汉语语料的合适位置生成跨语言语料，并通过弱监督学习扩展语料规模，最终通过神经概率语言模型学习得到汉泰词汇的跨语言分布表示，使在汉语上使用成熟的文本分析方法可以直接应用到泰语文本上，且在跨语言文本分析上的应用方法较为简单没有很复杂的消歧过程。实验通过文本相似度和文本分类验证取得了一定效果。我们下一步期望对神经概率语言模型进行改进(如增加隐藏层的层数等)来提高跨语言词汇分布表示的准确性，并进一步探讨跨语言词汇的分布特征向量表示维数对跨语言词汇分布表示的影响。

参 考 文 献

[1]Bengio S, Bengio Y. Taking on the curse of dimensionality in joint distributions using neural networks[J]. Neural Networks, IEEE Transactions on, 2000, 11(3): 550-557.

[2] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003, 3: 1137-1155.

[3]Collobert R, Weston J, Bottou L, et al. Natural Language Processing (almost) from Scratch[J]. Journal of Machine Learning Research, 2011, 12(1):2493-2537.

[4]Zeman D, Resnik P. Cross-Language Parser Adaptation between Related Languages[C]//IJCNLP. 2008: 35-42.

[5] Søgaard A. Data point selection for cross-language adaptation of dependency parsers[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics, 2011: 682-686.

[6] Ando R K, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data[J]. The Journal of Machine Learning Research, 2005, 6: 1817-1853.

[7] Prettenhofer P, Stein B. Cross-language text classification using structural correspondence learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 1118-1127.
[8]Steinberger R, Pouliquen B, Hagman J. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc[M]//Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2002: 415-424
[9] Wu L, Huang X, Guo Y, et al. FDU at TREC-9: CLIR, Filtering and QA tasks[J]. Proceedings of Text Retrieval Conference, 2000.
[10] Gao J, Nie J, Xun E, et al. Improving query translation for cross-language information retrieval using statistical models[J]. Sigir, 2001:96-104.
[11] Och F J, Ney H. Improved Statistical Alignment Models.[J]. Meeting of the Association for Computational Linguistics, 2000:440--447.
[12] Emami A, Jelinek F. A neural syntactic language model[J]. Machine learning, 2005, 60(1-3): 195-227.
[13] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5.
[14] Zeiler M D. ADADELTA: an adaptive learning rate method[J]. arXiv preprint arXiv:1212.5701, 2012.
[15]Shi L, Mihalcea R, Tian M. Cross language text classification by model translation and semi-supervised learning[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language ProcessingAssociation for Computational Linguistics, 2010:1057-106

作者一张金鹏（1989—），男，硕士研究生，主要研究领域为自然语言处理与嵌入式系统研究。
Email:939127870@qq.com

作者二周兰江（1964—），通讯作者，男，硕士生导师，副教授，主要研究领域为自然语言处理与嵌入式系统研究。
Email: 915090822@qq.com

作者三线岩团(1981—)，男，讲师，主要研究方向为信息检索、自然语言处理，E-mail: 195426286@qq.com

附论文修改说明：
带数字编号的为专家意见列表，意见下为相应修改说明
专家一意见：
1. 文章英文题目建议做修改，明显是病句。
已按要求对英文题目进行修改
2. 摘要要短小精炼，需缩减。
摘要经过缩减至 5 行
3. 在弱监督学习过程中引入人工判断，使得框架不能自动化运行，效率较低。

目前没有较好的方法通过知识逻辑的角度对实例进行校验，仍需一定程度的人工参与。

4. 实验中，“人工判定相似程度高于 95%”太过主观，削弱了实验的说服力。

汉泰维基百科中双语语料篇幅参差不齐，请语言学家进行判定双语文本的相似程度，提高说服力。

5. 文章学到的词分布式表示，当推广到文本相似度计算时，作者没有说明 tf-idf 算法是如何利用词向量的，审稿人也很迷惑，tf-idf 是基于词频的词袋方 法，文本的词频向量和分布式表示向量怎么可以结合到一起用？词汇的分布式表示和文本的分布式表示是两个概念，不能混淆，作者需要说明词向量是怎么表达文本 向量。

已经在实验章节 4.1 单独说明如何通过 tf-idf 算法结合词向量计算文本相似度。

5. 表 1 中的实验结果并不很好。

已经在扩大语料规模情况下多次重新实验，由于汉语与泰语的差异性和方法局限性效果有限。

6. 表 2 中的实验不具有合理性，跨语言的词语分布式表示为什么要和机器翻译结果做文本分类的比较？ 一方面，现阶段机器翻译的效果还有局限，翻译效果不好；另一方面，跨语言文本分类是个伪命题，没有价值。

   增加了与主流的跨语言文本分类方法（模型翻译、模型翻译与半监督适应结合）的对比实验

专家二意见：

1.However, the overall the novelty is limited. Besides, the experimental results are not compared with any strong baselines.：

增加了与主流的跨语言文本分类方法（模型翻译、模型翻译与半监督适应结合）的对比实验

专家三意见：

1. The English title and abstract of the paper is poorly written. I suggest the authors have naive English speakers proof read it.

   已经对英文论文题目与摘要进行了修改

2. I would like to see the comparison with an existing research work, other than Google translate.

   增加了与主流的跨语言文本分类方法（模型翻译、模型翻译与半监督适应结合）的对比实验