文章编号:1007-1423(2021)21-0008-07 DOI:10.3969/j.issn.1007-1423.2021.21.002

# 基于深度学习的双语词典构建方法研究

李祥祥,石刚

(新疆大学信息科学与工程学院,乌鲁木齐 830046)

**摘要:**
双语词典的构建方法一直是人们研究的重要方向。随着近些年理论水平和硬件性能的不断发展,基于深度学习神经网络的算法,在各个研究领域都取得了很大的突破。本文利用爬虫技术从网络上爬取汉英双语语料,经过分词、去停用词、词形还原等处理后,通过神经网络训练汉英双语语料,得到双语词向量,进而构建出汉英双语词典。实验结果表明,该方法在构建双语词典方面具有良好的效果。

**关键词:**
深度学习;汉英双语词典;双语词向量

## 0 引言

随着科技的发展,特别是互联网技术的普及,人们之间的距离越来越近,真正做到了“天涯若比邻”。远隔天涯的人们可以很方便地通过互联网交谈,了解对方的近况。伴随着全球化的不断迈进,世界各国之间的政治、经济、文化交流与融合也越来越密切,各国民众对于互相交流的意愿也越来越强烈,但是语言不通是极大地障碍,怎么让说不同语言的人无障碍的交流是自然语言处理的初衷。

双语词典是一种基本的资源,其在跨语言自然语言的很多工作中都起到了非常重要的作用[1]。传统的人工构造双语词典的方法有很多弊端,如人工成本较高、时效性不好且语言和领域的覆盖度方面不理想等。随个各个国家、各个民族的交流日益密切,每种语言都在不断的出现新的词汇,一个不能实时更新的双语词典很难满足跨语言自然语言处理的任务,如何更加实时且高效地构建双语词典成为了研究人员的一个重要课题。许多研究人员开始着手研究如何利用深度学习的方法自动完成双语词典的构建。

本文基于此背景,研究出了一种基于深度学习神经网络的双语词典构建方法。通过网络爬虫技术爬取汉英双语新闻网站、汉英翻译网站的数据,作为本文所用汉英语料。将汉英语料中语句分割成一个个词语(分词处理);分别设计了汉语和英语的停用词表,用于去除汉英语料中出现频率很高但没有实际含义的功能词(介词、冠词、代词等);最后使用神经网络模型训练汉英语料得到双语词向量,根据词向量的关联性得到汉英对照翻译词对。得到的双语对照翻译结果可作为专业编纂人员的参考和补充,避免从零开始工作,减少大量人工成本和时间成本,加快双语词典的构建速度。

## 1 双语词典构建方法

### 1.1 语料库

使用计算机算法构建双语词典主要有两种方式:

(1)基于平行语料库。平行语料库中的语句都是严格对照翻译的[1],利用这些严格互译的语句,可以很方便地获取优质,高效的双语词典。只是最大的问题是创建平行语料库比较耗时,并且双语词典覆盖的范围只在少数领域和少数语言中,很大程度上限制了这种方法的大范围使用[2]。

(2)基于可比语料库。可比语料库中的语句不是一一对照翻译的,但是互相翻译的词语在语料库中的位置是十分相近的。使用可比语料库进行双语词典构

建[3-4,8]的方式是从20世纪90年代末开始的发展起来的，由Fung和Rapp两人于同一个时间段提出这种观点[5-6]。这种方式的理论基础是：互译的词语通常出现在相似的上下文语境中。首先计算两种语言的词向量，然后计算词向量之间的距离，两个词向量的距离越近，对应词的相似度越高，互译的程度就越高，从而完成了互译词语的提取。

基于可比预料库的方法可以快捷且高效地构建双语词典，并且构建出的词典其语言和领域的覆盖都比较广。本文正是利用此方法的优点进行汉英双语词典的构建。通过网络爬虫技术爬取包含汉英双语的新闻、翻译类网站的公开数据，包含4000000个语句和段落，数据共计327MB，构成本文研究使用的汉英双语语料库。

### 1.2 构建方法

目前利用可比预料库构建双语词典的方法主要有：

（1）两种语言各自训练，互相转换。Mikolov[9]通过实验证明了，源语言的词向量能够很好地转变为目标语言的词向量[7,11]。该文章使用英语和西班牙语两种语言，通过训练分别得到对应的两种词向量空间E和S，从英语中取出one、two、three、four、five五个词语，从西班牙语中取出语义相同的五个词语uno、dos、tres、cuatro、cinco,将他们分别使用PCA（Principal Component Analysis，主成分分析）降维，通过降维解决维数灾难问题，并且在降维的同时让信息的损失最小，并在二维空间中描出来，发现他们在对应的词向量空间中的相对位置差不多，这说明不同语言在其对应的向量空间中的结构具有相似性。将源语言词向量转化到对应的目标语言词向量空间后，对照翻译的每两个词语的相对位置相近。

（2）两种语言各自训练，然后转换到第三方词向量空间。Faruqui[10]使用子词典将分别训练好的两种语言的词向量同时转化到第三方词向量空间中，由此得到的词向量空间中，互译的两种语言词语对会出现在相近的位置上。

（3）两种语言在同一词向量空间一起训练。代表工作是Stephan Gouws[12]将两种语言的语料混合在一起进行训练，得到在同一向量空间中的双语词向量。训练时的语料经过处理，具有互译性质的词语对通常出现在上下文中，同时训练出的具有互译性质词语的词向量通常具有很高的相似性，可以认定他们是互译的词语。

本文参考第三种方式，利用神经网络模型将汉英语料一起进行训练，得到同一向量空间中的双语词向量，通过计算词向量之间的关联性，可同时得到欲翻译词语的近义词和另一种语言的翻译结果。根据翻译结果完成双语词典的构建。

## 2 双语词典构建技术研究

在自然语言处理的应用中，需要将平时说的话以一种计算机能识别的方式输入，经常进行的操作是将语言数字化，词向量就是将语言中的词进行数学化的一种方式。

词向量大致有两种表示方式，第一种方式是One-hot-Representation,这种方式将一个词表示为一个很长的由很多个0和一个1组成的词向量，1所在的位置代表了唯一的一个词语。例如，“中国”的词向量可以表示为[0 0 0 1 0 0 0 0 0 0...]，“中华”的词向量可以表示为[0 0 0 0 0 0 0 1 0 0 ...]。用One-hot-Representation表示词向量的方法采用了稀疏方式存储，十分简洁，给每一个词语配置一个唯一的编号。这种表示方式配合上最大熵、SVM、CRF等算法可以很好地完成自然语言处理领域的各种主流任务。然而这种方式也有一些缺陷，它很容易受到维数灾难的困扰，这是一种随着数据维数的增加，计算量呈指数倍增长的现象；同时该方法无法准确描述词和词之间的语义关系，如刚才举例的“中国”和“中华”，虽然词义是相似的，但是One-hot-Representation并不能描述出他们的相似性。第二种方式是Distributed Representation,使用此种方法可以将词语表示为一种维数很低的实数向量，一般表示成：[-8.70686, -1.8112, -2.19500, -1.15472, -2.82405, ...]。这种词向量表示方法可以很好地刻画语义相似的两个词语的相似度，理想情况下“中国”和“中华”词向量的相似度会十分高，会远远大于“中国”和“钢笔”的相似度。

词向量实际上是在对语言进行建模的同时获取到的一种词语在向量空间中的表示，是语言建模时的一种副产物。语言建模就是判断一个语句是否是正常人说出来的，这种建模有很重要的意义，比方说机器翻译和语音识别，当得到若干个候选语句后，使用语言模

型，可以选择出最佳的结果。给定一个句子 $S$，由 $t$ 个词语 $W_1, W_2, W_3, \cdots, W_t$ 组成，这个句子是自然语言的概率为 P(S)，$P(W_1)$ 代表词语 $W_1$ 出现的概率，$P(W_2|W_1)$ 代表在有词语 $W_1$ 存在的情况下 $W_2$ 出现的条件概率。因此词向量的计算等同于计算 $P(W_i|Context_i)$ 问题，计算在 $Context_i$（词语 $W_1$ 至 $W_{i-1}$）存在的情况下，$W_i$ 出现的概率。使用数学表达式表示语言建模，如公式（1）所示：

$$P(S)=P(W_1,W_2,W_3,\cdots,W_t)$$
$$=P(W_1)*P(W_2|W_1)*P(W_3|W_1,W_2)*\cdots*P(W_t|W_1,W_2,\cdots,W_{t-1})$$
$$=\prod_{i=1}^{t}P(W_i|Context_i) \quad (1)$$

Google 曾公布一种名为 Word2Vec 的模型[13]。它因为可以简单、高效得到词向量而引起人们的关注。

Word2Vec 的作用就是将平时说话时所用的词语转化成计算机可以识别的词向量。它本质上是一种单词聚类的方法，是实现单词语义推测、句子情感分析等目的的一种手段。选取训练后词向量的任意 3 个维度，放入坐标系中，会发现语义相似的词语在空间坐标中的位置十分接近，而语义无关的词语之间则距离较远，这种性质可以很好地描述单词之间的相似性。Word2Vec 采用三层的神经网络，即输入层、隐层和输出层，其核心技术是根据词频使用 Huffman 编码，使得所有词频相似的词语在隐藏层激活的内容基本一致，出现频率越高的词语，它们激活的隐藏层数目越少，这样有效地降低了计算的复杂度。

Word2Vec 分为两种类型的训练方法：CBOW 模型和 Skip-Gram 模型。CBOW 的是利用上下文来预测中心词语的出现的概率，而 Skip-Gram 与之相反，利用中心词语来预测上下文的概率。在训练的最开始，给每个单词都设置一个随机的 K 维 One-hot-Representation 向量作为输入数据，经过 CBOW 或者 Skip-Gram 模型进行训练后，可以得到每个单词的最合理的 n 维 Distributed Representation 向量。

CBOW 最主要的思想是利用上下文来猜测中心词语出现的概率。（如图 1 所示）设定 $w(t)$ 为要预测的词语，设置窗口大小为 2，输入层输入的数据为 $w(t)$ 上下的两个词语 $w(t-2)$、$w(t-1)$、$w(t+1)$、$w(t+2)$ 四个词语，隐层的 SUM 为输入层输入数据的累加和，输出层对应了一棵 Huffman 树。

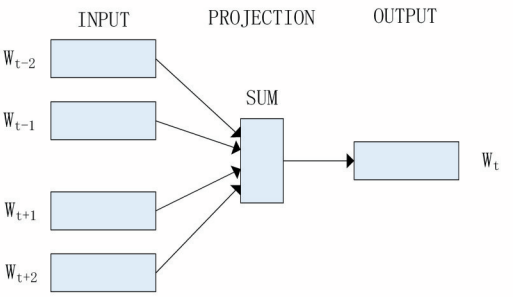


图 1 CBOW 模型

图 1 CBOW 模型设窗口长度为 2，中心词如图 3 所示移动，蓝色区域为中心词，灰色区域为中心词的上下文。假设输入词为 input，预测的词为 label，每一次训练的样本为（input,label）。对于 CBOW 模型，每移动一次中心词，可以产生一个训练样本，如图 2 所示，中心词移动了四次，产生了四个样本，分别为（[great,was], nothing），（[nothing,was,ever],great），（[nothing,great,ever, achieved],was），（[great,was,achieved, without], ever）。本文使用 CBOW 模型，通过训练得到词语的上下文信息，根据两种语言词语共同出现的概率（共现度）判断是否为对照翻译的词语。

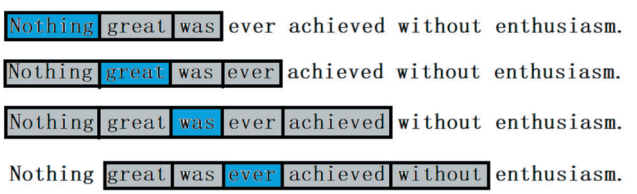


图 2 训练样本

## 3 实验过程

首先通过网络爬虫技术收集英语及汉语语料各约二百万条语句，这些语句包含日常生活的方方面面。对语料库进行分词处理，将语句段落分解成一个个词语。英语的分词较为简单，因为英语的词与词之间是靠空格隔开的，进行英语分词的时候只需要根据空格和英语符号进行分词，再去除英语的停用词就行。而中文的语句是以字为单位，一个个字连起来组成了有意义的词语和句子。例如，英语中的句子“I am a fresh graduate of Xinjiang University”，翻译成中文就是：“我是新疆大学的应届毕业生”。计算机可以很轻易地通过

空格就分析出 graduate 是一个单词，但是却很难理解“毕”、“业”、“生”三个字组合起来才是一个词语。把汉语的语句切分成人类可以识别出的有意义的词语，就是中文分词。“我是新疆大学的应届毕业生”这句话的正确分词结果是：“我”“是”“新疆大学”“的”“应届”“毕业生”。

### 3.1 中文分词处理

为了进行中文的分词，本文使用了优秀的中文分词工具“jieba 分词”，结巴分词支持多种分词方式，分词效果十分理想。其对汉语中的所有词语构建了一个词典，每一个词语都有三个重要属性：词语、词频和词性。当对一个文本进行分词的时候，会根据词典中词语的词频和词性进行分词，词频越高被判定为一个词语的几率越高。也可以根据需要自己构建词典，以满足特定的需要。对二百万条汉语语料进行分词处理之后，需要进行去停用词操作。停用词(Stop Words)是在进行文本检索或者搜索引擎索引的时候为了提高效率而进行忽略操作的某些字、符号或词语。本文为了完成汉语的去停用词的操作，设计了一个拥有 1893 个停用词的汉语停用词表，基本满足需要。分词之后的大多数词语都是没有实际意义的停用词，停用词跟任何词语的共现度都很高，对基于词语共现度进行词向量生成的 Word2Vec 来说，会严重影响词向量的准确率，所以去除停用词必不可少。

### 3.2 英文分词处理

进行英文分词虽然比较简单，但是还是有很多值得注意的地方：

(1)需要将所有的英语语句中的大写字母转换为小写，因为具有大写字母的词语虽然在词义上和小写的相同，但是在训练模型中会被认成两种词语，影响训练的效果。

(2)本文使用的英语语料由语句和段落组成，会出现很多不同形态的词语，例如名词的单复数形式，动词的进行时、过去式、被动形式、形容词的比较级等。这些词语的语义相同，形态不同，会被当成不同的词语进行处理，但中文中并没有形态的困扰，不同形态的英文单词只有一个形态的中文翻译结果。需要对不同形态的词语进行词形还原处理，将所有的词语转换为原形，提高训练的准确率。本文使用“NLTK”作为词形还原处理工具，全称为 Natural Language Toolkit。是一套基于 Python 的自然语言处理工具集。NLTK 进行词形还原的时候根据不同的词语类型进行操作，例如名词、动词、形容词等。词形还原处理示例如表 1 所示。

表 1 词形还原表

| 词语 | moderately | nurses | reading | goods | better |
|---|---|---|---|---|---|
| 名词词形还原 | moderately | nurse | reading | goods | better |
| 动词词形还原 | moderately | nurse | read | goods | better |
| 形容词词形还原 | moderately | nurses | reading | good | good |

不同词性词语经过对应的词形还原后，可以得到符合要求的词语。NLTK 具有词性标注功能，可以识别不同单词的词性，再根据词性调用对应的接口将词语处理为原形。

(3)英文分词处理。利用每个词语之间的空格和符号将每个词语分割出来，然后设计一款英语停用词表，去除分词后英语语料中的停用词。英语停用词包括冠词(a,an,the)、介词(in,on,from,above,behind)、连词(and,but,before)、感叹词(oh,well)、代词(who,she,you,it)还有英语的标点符号等。与汉语情况相同，这些停用词在每一句英语句子中几乎都有，却并没有有效的含义，会严重影响词向量和共现词语的准确率，必须要去除掉。本文设计的英语停用词表包含 891 个停用词，在实际的使用中表现良好。

### 3.3 合并双语语料

将处理好的汉语和英语的分词结果交叉合并为一个语料库，此语料库中都是本文需要的汉语及英语词语。合并后的语料库如图 3 所示。


图3 双语分词语料库

### 3.4 训练与结果

本文使用 Word2Vec 中的 CBOW 模型进行训练。深度学习框架使用 TensorFlow，TensorFlow 由谷歌公司

研发，可以应用于语音识别，自然语言处理，计算机视觉等多项深度学习领域。提供 Python 接口，能够使多个 CPU 和 GPU 协同工作，具备很好的灵活性和可扩展性，支持异构设备的分布式计算，能够支持 CNN、RNN、LSTM 等算法，是一款相当优秀的深度学习框架。本文的词向量训练过程分为以下几步：

（1）读取经过分词和去停用词处理的汉英双语语料库，作为训练语料。该语料库中总共有 27495541 个词语。

（2）统计每一个词语的频率。取用频率最高的前 40 万个词语，对每一个词语设置一个索引。根据统计结果，出现频率在 2 以上的词语有 331090 个，频率最高词语的频率为 137890。如表 2 所示：频率代表的是每个词语出现的次数，个数代表的是大于这个频率的总共有多少个词语。

表 2 词语频率及出现个数

| 频率 | >2 | >100 | >500 | >1000 | >2000 | >5000 | >10000 |
|---|---|---|---|---|---|---|---|
| 个数 | 331090 | 27155 | 8366 | 4755 | 2500 | 825 | 271 |

（3）为 CBOW 模型创建一个批处理函数来量化训练样本的大小及范围。

（4）构造基于 TensorFlow 的 CBOW 的计算图，其核心是定义计算损失（loss）的公式以及计算中使用的优化方法。

（5）开启会话，运行构造好的计算图，对模型进行训练。

（6）在最后一次训练完后，将得到的模型保存起来，以待下一次的调用。

（7）根据训练好的双语词向量，使用 sklearn 的 TSNE 对词向量进行降维处理，对降维数据以图像形式直观地展示（如图 4 所示），其中语义相近（包括近义词和互译词）的词语相对位置较近。

（8）加载训练好的词向量模型，使用频率最高的三十万个单词作为输入，得到对应的翻译结果，并进行保存。当输入是汉语词语后，只保存对应的英语翻译词语，当输入的英语词语时，只保存对应的汉语翻译词语。之后使用一个标准的汉英词典与本文构建的词典进行对比，只保留互为翻译的汉英词对。得到的结果就是本文构建成功的汉英双语词典，如图 5 所示。

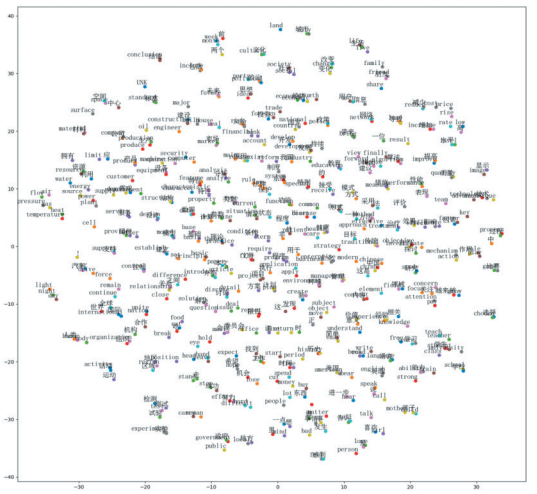

图 4 词向量降维后词语位置

```
研究: study, research,
system: 系统, 制度,
time: 时间,
方法: method,
method: 方法,
发展: development,
study: 研究, 课题,
工作: job,
中国: china,
分析: analysis,
development: 发展,
影响: influence, affect, impact, effect,
china: 中国,
系统: system,
process: 过程, 加工,
提供: provide, offer,
life: 生活,
```

图 5 双语词典

基于 Python 中的 pyqt5 模块做的一个图形操作界面，对其输入框输入想要翻译的词语（无论是汉语还是英语词语），可以得到对应的翻译结果（如图 6 所示）。

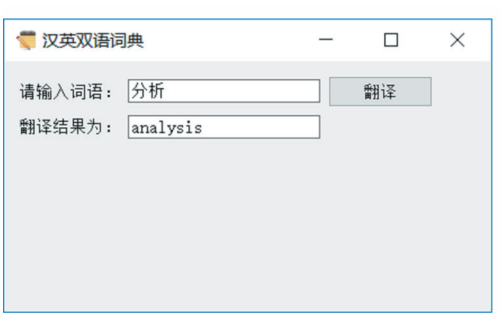


图 6 双语词典应用界面

## 4 词典准确率研究

### 4.1 已构建汉英双语词典结构研究

已构建的汉英双语词典中汉语和英语词语按照训练时词语出现频率的顺序由上至下排列，一个词语占一行空间，每一行的第一个词语是训练时的词语，后面跟随多个训练出来的翻译词语。如图 7 所示：第一行中的“UNK”代表频率排名 4 万名以后的词语。

```
UNK: 二重奏, 状, 加,
中: role, kldo, compital,
研究: study, research, thesis,
system: 系统, 控制系统, 制度, 体系, 管理系统, 实时, 子系统,
time: 时间, 花费, 票, 浪费, 花, 时光, 每次,
方法: method, computation,
method: 方法, 相结合, 法, 采用, 分析方法, 计算方法, 提出,
people: 全世界, 带给, 当风, 民众, 大众, 扑克,
发展: development, sustainable, promote, develop, globalization,
一种: adaptive, propose, algorithm, multi, overcome,
study: 研究, 研究成果, 学, 热点, 机理, 课题, 国内外,
工作: job, employee, employer,
时: permission, runtime, default, ashless,
中国: china, chinese, lu, cun, zhen, yuan,
分析: analysis, analyze, analyse,
development: 发展, 经济社会, 推动, 持续, 飞速发展,
说: speak, dare,
影响: influence, affect, impact, effect, factor, influential,
china: 中国, 电子地图, 村, 镇,
paper: 本文, 介绍, 论述, 文中, 原理,
```

图 7 汉英对照翻译词对

图 7 中每个词语训练后的翻译结果有多个，但是真实的翻译情况只包含其中的一个或者多个结果，可以看出频率排名较高的词语翻译结果比较优秀，误差不大。

### 4.2 标准英汉词典的结构研究

为了研究出使用词向量构建出的汉英双语词典的翻译结果的准确率，本文使用一个标准的严格互译的汉英词典与其进行对比，推算出其翻译的准确率。

本文选取的标准互译的汉英词典如图 8 所示。词典按照 a-z 的顺序依次排列，前半部分是 7880 个单词，后半部分是 359 个短语，共 8239 个词语。但由于本文使用的双语语料库中只包含单词，没有短语，所以真正能用的只有前半部分的 7880 个单词。

```
abandon 放任；狂热；遐弃；放弃
abandonment 抛弃；放纵
abbreviation 缩写；缩写词
abeyance 中止；停顿；归属待定；暂搁
abide 忍受；容忍；停留；遵守；持续；忍受；停留
ability 能力；能耐；才能
able 能；[经管]有能力的；能干的；(Able)人名；(伊朗)阿布勒；(英)埃布尔
abnormal 反常的；不规则的；变态的
aboard 在飞机上；[船]在船上；在火车上；在…上
abolish 废除；废止；取消；革除
abolition 废除；废止
abortion 流产；堕胎；小产；流产的胎儿；（计划等）失败；夭折
abortive 失败的；流产的；堕胎的
about 关于；大约；在附近的；四处走动的；在起作用的；大约；周围；到处；大致；粗枝大叶；不拘小节的人；(About)人名；(法)阿布
above 超过；在……上面；在……之上；在上面；在上文；上文的；上文
```

图 8 标准英汉词典

### 4.3 对比求得词典准确率

本文使用词向量方案构建的汉英双语词典与标准互译的汉英词典进行对比，得出如下结果（如图 9 所示）。

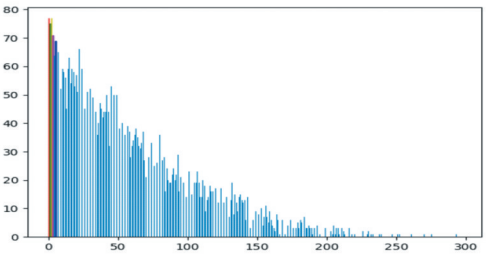


图 9 不同词频的翻译情况

对比后结果显示互相翻译的词语共有 5046 对。相较于标准汉英词典的 7880 个单词，已经比较符合预期。因为每一种词频的单词，其单词数目不一，于是本文根据词频的顺序排列，每 100 个单词，计算一次准确率。如图 8 所示，X 轴表示单词的相对顺序位置，X 轴的数值越大，单词的频率越小，轴上的数字表示条形的多少，每 100 个单词显示一个条形；Y 轴表示这 100 个单词中正确翻译的数目。由图可以看出训练时出现频率越高的词语，其训练的翻译结果就越准确。

## 5 结语

本文提出的基于深度学习构建双语词典的方案，可以减少人工成本，提高构建的速度。无论是构建基于新语言的双语词典，还是提炼网络和生活中出现的新鲜词语，都可以在使用收集语料的基础上，使用深度学习的方法自动完成双语词典的初步构建。构建出的词典与实际翻译结果会有一定的误差，但可以作为一种参考或双语词典的初始版本，经过专业编纂人员的校对和补充后，完成准确、可信的双语词典最终版本。针对网络和生活中新出现的词语，只需要利用网络爬虫技术爬取近些年互联网中的语料信息，使用本文方案便可以得出对应的英语翻译结果和近义词，可帮助专业人员研究新词的含义和演变过程。

参考文献：
[1]FUNG P. Finding terminology translation from non-parallel corpora[C]. 5ht Workshop on Very Large Corpora，1997：192-202.
[2]TAMURA A，WATANABE T，SUMITA E. Bilingual lexicon extraction from comparable corpora using label propagation[C]. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2012：24-36.
[3]陈鹏. 基于词向量与可比语料库的双语词典抽取算法研究[D]. 华中师范大学，2017.
[4]李舰，李波，陈鹏，等. 基于可比语料库的双语词典抽取方法比较研究[J]. 小型微型计算机系统，2017，38(7)：1554-1561.
[5]FUNG P. Compiling bilingual lexicon entries from a non-parallel engligsh-chinese corpus[C]. The Workshop on Very Large Corpora，1995：173-183.
[6]RAPP R. Identifying word translations in non-parallel texts[C]. Meeting of the Association for Computational Linguistics，26-30 June 1995，Mit，Cambridge，Massachusetts，Usa，Proceedings. DBLP，1995：320-322.
[7]GARERA N，CALLISON-BURCH C，YAROWSKY D. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences[M]，2009：129-137.
[8]YU K，TSUJII J. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity[C]. Human Language Technologies：the 2009 Conference of the North American Chapter of the Association for Computational Linguistics，Companion Volume：Short Papers. Association for Computational Linguistics，2009：121-124.
[9]MIKOLOV T，LE Q V，SUTSKEVER I. Exploiting similarities among languages for machine translation[J]. Computer Science，2013.
[10]FARUQUI M，DYER C. Improving vector space word representations using multilingual correlation[J]，2014.
[11]RADINSKY K，AGICHTEIN E，GABRILOVICH E，et al. A word at a time：computing word relatedness using temporal semantic analysis[C]. International Conference on World Wide Web. ACM，2011：337-346.
[12]GOUWS S，BENGIO Y，CORRADO G. BilBOWA：fast bilingual distributed representations without word alignments[J]. Eprint Arxiv，2014：748-756.
[13]MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[J]. Computer Science，2013.

作者简介：
李祥祥(1995-)，男，安徽人，硕士，研究方向为自然语言处理、计算机视觉
通信作者：石刚(1986-)，男，新疆乌鲁木齐人，副教授，硕士生导师，研究方向为系统软件、大数据处理，E-mail：shigang@xju.edu.cn
收稿日期：2021-03-04　　修稿日期：2021-03-27

# Research on the Construction Method of Bilingual Dictionary Based on Deep Learning

LI Xiangxiang, SHI Gang

(College of Information Science and Engineering, Xinjiang University, Urumqi 830046)

Abstract:
The construction of bilingual dictionaries has always been an important research direction. With the continuous development of theoretical level and hardware performance in recent years, the algorithm based on deep learning neural network has made great breakthroughs in various research fields. In this paper, crawler technology is used to crawl Chinese English bilingual corpus from the network. After word segmentation, stop words removal and word form reduction, neural network is used to train Chinese English bilingual corpus to get bilingual word vector, and then a Chinese English bilingual dictionary is constructed. The experimental results show that this method has a good effect in the construction of bilingual dictionary.
Keywords:
Deep Learning; Chinese-English Bilingual Dictionary; Bilingual Word Vector