计算机算法类资料的中英文智能翻译

陈家乐,张艳玲
(广州大学 计算机科学与网络工程学院,广东 广州 510006)

摘 要:当前互联网免费可用的在线翻译系统均是使用通用语料训练出来的神经机器翻译模型,在通用语义环境下翻译出色,而在特定的垂直领域(如计算机专业领域)中,由于训练文本和模型训练算法缺乏针对性,导致翻译结果出现专业词汇错漏,文本晦涩难懂。因此,实现特定垂直领域的自动化机器翻译的需求越来越大。通过网络爬虫获取计算机算法类相关的英汉双语例句,基于Word2Vec算法生成含有上下文信息的词向量,将词向量嵌入到Google开源GNMT模型训练英汉翻译模型,基于训练模型实现简易翻译软件。通过对照实验,探究Word2Vec算法中词向量长度对计算词汇间文本相似度的影响和对GNMT训练效果的影响,以及GNMT超参数中的隐藏层单元数num_unit、批尺寸batch_size对训练效果的影响。综合实验结果训练最佳的英汉翻译模型。

关键词:机器翻译;Word2Vec算法;词向量;文本相似度;GNMT
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2021)07-0176-06
doi:10.3969/j.issn.1673-629X.2021.06.029

English-Chinese Intelligent Translation of Computer Algorithm Corpus

CHEN Jia-le,ZHANG Yan-ling
(Faculty of Computer Science and Network Engineering,Guangzhou University,Guangzhou 510006,China)

Abstract:At present,the free and available online translation systems on the Internet are all neural machine translation models trained by general corpus,which are excellent in the general semantic environment. However,in the specific vertical field (such as computer professional field),due to lack of pertinence of training text and model training algorithm,the translation results appear professional vocabulary errors and omissions,and the text is obscure. Therefore,the demand to achieve an automated machine translation in a specific field becomes bigger and bigger. The English-Chinese bilingual example sentences related to the computer algorithm are obtained by web crawler,and the word vector with context information based on Word2Vec algorithm is generated and embedded into Google open-source GNMT model to train English-Chinese translation model. On the basis,a simple translation software is implemented. Through a comparative experiment,we explore the influence of word vector length on the calculation of text similarity between words and the training effect of GNMT in Word2Vec algorithm,as well as the influence of the number of hidden layer units and batch size in GNMT super parameters on the training effect,training the best English-Chinese translation model based on the experimental results.

Key words:machine translation;Word2Vec algorithm;word vector;text similarity;GNMT

0 引 言

自2014年以后,端到端的神经机器翻译(end to end neural machine translation)质量较统计机器翻译有了显著提升[1],端到端的训练使得深度学习方法区别于传统机器学习方法,成为了自然语言处理的强大工具[2]。Google、有道、百度等商用在线机器翻译系统核心技术均由统计机器翻译转型为神经机器翻译。端到端的编码器-解码器(Encoder-Decoder)[3]模型结构成为神经机器翻译的主流模型[4]。Google首次提出基于Encoder-Decoder结构的seq2seq[5](sequence to sequence)模型。其基本工作原理就是编码器将输入序列转换为中间向量C,解码器将中间向量C转换为输出序列。seq2seq在编码器和解码器中均使用循环神经网络(recurrent neural networks,RNN),理论上RNN可以解决长句远距离信息依赖的问题[6],但在实际应用时反向传播过程中存在梯度爆炸和梯度消失的问题。梯度爆炸会导致模型无法收敛;梯度消失会导致模型捕捉不到长距离项的依赖信息。梯度爆炸一般可以使用梯度裁剪或权重正则化处理。梯度消失目前最好的处理方法是使用门控单元来构建RNN。其中,

收稿日期:2020-07-12 **修回日期**:2020-11-13
基金项目:2018年教育部第二批产学合作协同育人项目(201802093015)
作者简介:陈家乐(1998-),男,研究方向为自然语言处理;通讯作者:张艳玲(1970-),女,博士,副教授,硕导,研究方向为人工智能及其应用。

应用最广的门控单元是长短时记忆(long short-term memory,LSTM)和门控循环单元(gated recurrent unit,GRU)[4]。虽然应用门控单元使得 RNN 能有效捕捉到长距离项的依赖信息,但是由于句子中任意单词对生成某个目标单词的影响占比是相同的,所以即使采用了门控单元,seq2seq 模型依然在长句的翻译质量上显著下降。为了优化长句的翻译质量,Bahdanau 等人[7]于 2014 年首次将注意力机制应用于 NLP 领域,翻译效果有了进一步提升。谷歌团队[8]抛弃 RNN 和 CNN 等网络结构,仅仅采用注意力机制进行机器翻译,在翻译质量上取得了显著的效果,因此注意力机制成为了神经机器翻译中的研究热点。但是,由于通用翻译接口对垂直领域缺乏针对性,同一个词汇在不同的语义环境下有不同的翻译结果,而通用翻译则无法识别该词汇所在的语义环境,从而使得翻译效果不佳。并且垂直领域下的专业词汇繁多且复杂,若没有对应的词汇的语料训练,会让最终的训练模型对含有该词汇的句子没有好的翻译效果。这就是现阶段对垂直领域的语句翻译效果不好的原因。所以针对某一领域实现翻译成为了当下重要的研究方向。

该文将收集与计算机算法类相关的中英双语例句文本,利用 Word2Vec 算法生成词向量,将词向量嵌入 GNMT[9]训练带有注意力机制的 LSTM seq2seq 的中英翻译模型。以此来优化计算机算法类语料的翻译效果,为此后垂直领域的神经机器翻译提供一个可行的优化思路。

1　数据获取

数据样本的好坏很大程度上决定了模型的训练效果,要实现计算机算法类资料的中英翻译,就需要获取与计算机算法类相关的中英双语例句。现在各大在线翻译网站能够检索特定关键词的中英双语例句,但是大多数的双语例句没有分门别类。为了得到最符合计算机算法类的中英双语例句,需要收集计算机算法类相关度高的关键词。

该文通过中文书籍《算法导论第三版》(Introduction to Algorithms Third Edition)和计算机算法题网站力扣(LeetCode)收集关键词。收集了 1 618 个关键词在百度翻译、有道翻译、知网例句进行检索并爬取,最终获取 98 120 条中英双语例句作为模型训练的数据基础。

2　数据预处理

获取的数据还不能直接用于数据训练,中文例句词汇与词汇之间并没有明确的分隔;英文例句虽然词汇间有天然的空格分割,但是标点符号与词汇间仍然有连接,因此也需要进行分词处理。分词处理完成后还需要获取训练所需的数据集和词汇表。

2.1　英文文本处理

英文存在大小写的区别,而大写的写法和小写的写法指的是同一个单词,例如“The”与“the”。经过小写化处理后,获取词集时能够减少大量的重复单词,从而降低训练的成本,一定程度上优化训练效果。还需要对英文文本分词,将标点符号与词汇间进行分隔。

2.2　中文文本分词

中文分词正确率会大大影响模型训练的效果,因为分词阶段的错误在翻译过程中将会被“放大”,放大的倍数约等于句子的平均长度[10]。jieba 库是一个简单易用的汉语自然语言处理分词库,通过在全切分所得的所有结果中求某个切分方案 S,使得 P(S)最大的概率进行分词。jieba 分词的算法流程为:(1)基于特定词汇表构建字典树,实现高效的词图扫描;(2)基于字典树生成句子中汉字所有可能成词情况所构成的有向无环图(DAG);(3)采用了动态规划寻找最大概率路径,找出基于词频的最大切分组合;(4)对于未登录词,使用了 Viterbi 算法并采用了基于汉字成词能力的 HMM 模型[11]。jieba 分词有三个分词模式,分别是精确模式、全模式、搜索引擎模式。其中精确模式适合用于自然语言处理。

中文一词多意的情况非常多,同一个句子在不同的语义环境下有不同的分词方案,因此如果不制作对应的分词词典,会大大增加分词出错的概率。制作分词词典步骤如下:(1)对中文文本进行默认分词;(2)人工将分词出错的词汇添加到词典中。

2.3　数据集分割

在模型训练之前,首先要划分训练集、验证集、测试集。其中训练集用于模型训练,验证集和测试集用于衡量模型训练的效果。该文采用的分割方式是生成长度为 0 ~ $N-1$(N 为句子总数)的乱序序列,将乱序序列按 14∶3∶3 的比例进行分割,分别对应训练集、验证集、测试集。再根据分割后的乱序序列取出对应下标值的句子存入到对应的文件当中。

2.4　基于训练集的中英文词集提取

首先,定义三个标签:<unk>,<s>,</s>;其中<unk>表示未定义词汇,<s>表示语句的开头,<\s>表示语句的结束。由于文本已经经过分词处理,故只需按行读取句子,根据空格进行分割就能够获取到词汇,并且按照词汇的频率由高到底排序。

3　词向量训练

3.1　文本表示方法

文本表示方法一直是自然语言处理研究范畴中的

一个热点问题,总体来讲主要分为两大类:独热编码和分布式表示。

独热编码(one-hot representation)又称为一位有效编码,这种编码格式是建立一个全局完备的字典,但在计算上面临着两个问题,一个是这种表示方法的向量维度是字典的大小,而字典中的词汇数目往往很大,从而在计算时避免不了维数灾难的问题,给计算机带来极大的负担;另一个是这种表示只包含词汇在字典中的索引和词频信息,未考虑词的上下文信息,无法从向量上判断两个词汇是否相似,不能为后续的模型训练提供更多有用的信息。

分布式表示(distributed representation)是一种稠密、低维的实值向量表示,由 Hinton[12] 在 1986 年提出,能够有效克服独热编码的缺点。每个维度表示单词的不同句法和语义特征。词向量是一种词汇的分布式表示形式,通过对文本语料库进行训练,将每个词用 N 维的实值向量表示,向量可以看作空间上的一条线,通过计算向量之间形成的角度,就可以判断两个单词之间的相似度。其中 Word2Vec 是一个可以快速训练词向量的算法。

3.2 Word2Vec 算法

Word2Vec 算法是 Tomas Mikolov 带领的研究团队发明的[13]。其基本思想是利用上下文信息,即使用与当前词相邻的若干个词,来生成当前词的特征向量。其中包含了跳字模型(Skip-gram)和连续词袋模型(CBOW)两种训练模型,CBOW 是通过上下文来预测当前词,而 Skip-gram 则相反,它是通过当前词来预测上下文。同时,Word2Vec 提供了两套优化方法来提高词向量的训练效率,分别是 Hierachy Softmax 和 Negative Sampling[14]。通过将训练模型与优化方法进行组合可以得到 4 种训练词向量的架构。

3.3 基于 Word2Vec 算法训练词向量

通过 Python 中 gensim 库里封装的 Word2Vec 进行词向量的训练。主要的训练参数说明如表 1 所示。

表 1 Word2Vec 训练参数说明

参数	说明
sentences	需要训练的语料集
sg	设置训练算法,默认为 0 使用 CBOW 算法;为 1 采用 Skip-gram 算法
size	生成词向量的维度
window	当前词与预测词在一个句子中的最大距离
min_count	词频少于 min_count 的词汇将会被丢弃
workers	训练时的并行数
hs	设置优化方法,默认为 0 使用 Negative Sampling;为 1 采用 Hierachy Softmax

使用训练集作为训练语料,主要的调试参数为词向量的维度,固定参数:sg=1,min_count=1,work-ers=4,hs=1。其中参数 window 若设置得过小,将无法捕捉句子中较长距离的依赖信息,若设置得过大,将捕捉到过多的无效依赖,从而降低了有效依赖的权重,因此将固定 window 参数为 8。通过将词向量的维度设置为 1,2,4,8,16,32,64,128,256,512 进行训练,从而探究词向量维度对计算词汇间文本相似度的影响。对于中文文本,将使用“算法”,“复杂度”,“排序”对模型进行测试。对于英文文本,将使用“algorithm”,“complexity”,“sort”对模型进行测试。

3.4 训练结果

对不同词向量维度模型进行相似度测试后,通过对前十的相似度词汇的相似度取平均数画出相似度变化折线,如图 1 所示。

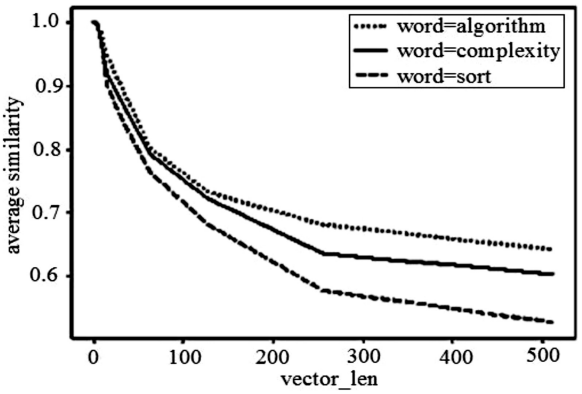


图 1 平均相似度变化折线

由图 1 可以得出:同一个单词,词向量维度越高,计算所得的单词相似度越低。通过对测试结果中的实际相似单词进行分析,可以看出:当词向量维度为 1 时,无论使用哪个词汇进行测试的结果都是一样的;当词向量维度过低时,测试结果含有实际意义的词汇较少;随着词向量维度的增加,相似度测试结果具有实际意义的词汇渐渐变多。因此,剔除维度为 1,2,4,8,16 的词向量,保留维度为 32,64,128,256,512 的词向量进行下一步的翻译模型训练。

4 基于 GNMT 训练翻译模型

神经机器翻译(neural machine translation,NMT)存在致命的缺点,即计算成本非常昂贵,并且大多数 NMT 系统对罕见词的处理效果不好且在处理长句有翻漏的现象。谷歌发布的 GNMT(Google’s neural machine translation)解决了上述问题,翻译误差平均降低了 60%。在推理过程中采用低精度算法以及 TPU 进行计算还可以解决翻译速度问题。为了更好地处理罕见词,将罕见词拆分为常见子词单元分别进行处理。为了减少长句翻漏的现象,在波束搜索中使用长度规范化过程和覆盖度惩罚机制[3]。使用 Google 在

GitHub 上开源的代码训练带有注意力机制的两层 LSTM seq2seq 模型,首先探究词向量对模型训练的影响,再探究隐藏单元数 num_unit 以及批尺寸 batch_size 对模型训练的影响。最终选取最佳的参数进行训练。具体实验步骤为:

(1)使用词向量维度为 32,64,128,256,512 进行训练,选取效果最佳的词向量维度进入下一个实验;

(2)使用 num_unit 为 32,64,128,256,512 进行训练,选取效果最佳的 num_unit 进入下一个实验;

(3)使用 batch_size 为 4,32,64,128,192,256 进行训练,选取效果最佳的 batch_size;

(4)综合实验结果,选取最优参数进行训练,直到模型不再优化为止。

5　实验结果

模型训练的环境如表 2 所示。

表 2　模型训练环境

处理器	内存	显卡	操作系统	编程语言	Tensorflow 版本
AMD Ryzen 3600x @4.0 GHz	16 GB	NVIDIA GeForce GTX 1080Ti	Ubuntu 18.04	Python 3.6.0	Tensorflow-gpu 1.8.0

5.1　翻译指标 ppl 和 bleu

ppl 指的是困惑度(perplexity),是统计机器翻译中的评价指标,用于评判机器翻译的译文是不是一个合理的语句。它是通过对概率平均数取倒数计算获得,所以当模型的翻译结果越合理,困惑度越低。

bleu(bilingual evaluation understudy)[15]是由 IBM 于 2001 年提出的一种文本评估算法,用来评估机器翻译与专业人工翻译之间的接近程度,核心思想就是当机器翻译越接近人工翻译,bleu 分数越高,说明机器翻译与人工翻译之间越接近。

5.2　词向量维度对模型训练的影响

将使用词向量维度为 32,64,128,256,512 进行模型训练。基于 ppl 及 bleu 进行对照的折线变化如图 2 所示。

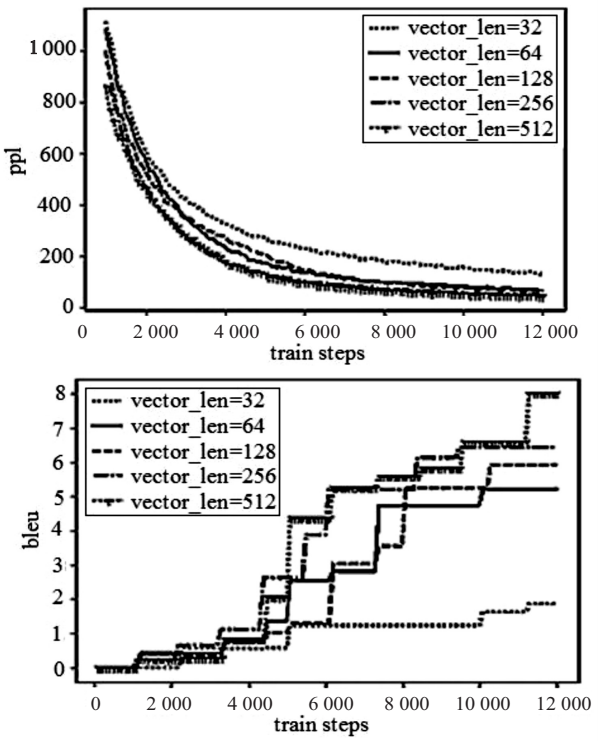


图 2　不同词向量维度模型的 ppl 及 bleu 对照折线变化

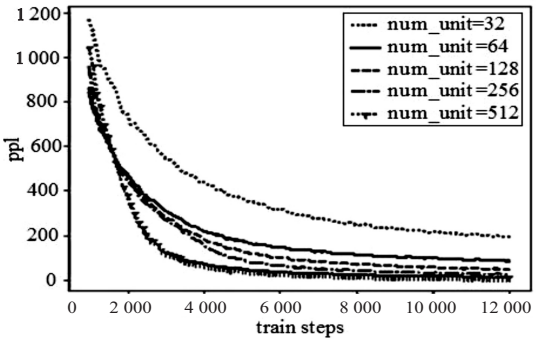
由图 2 可见,随着训练步数的增大,ppl 的变化越来越小,慢慢趋近于没有变化。随着词向量维度的增大,最终的 ppl 变小。随着训练步数的增大,每个模型的 bleu 值都呈上升趋势。在模型训练中期,词向量维度低的模型的 bleu 值超越了词向量维度高的模型的 bleu 值,但只是暂时的,随着训练步数的增加,最终词向量维度高的模型 bleu 值变大。

综上所述,在词向量维度取值为 32,64,128,256,512 时,词向量维度为 512 时的模型训练效果最优。并根据图像显示,继续增大词向量维度,模型有进一步优化的可能。

5.3　超参数 num_unit 对模型训练的影响

超参数 num_unit 指的是隐藏层单元数,过小的 num_unit 会使神经网络的表达能力差,从而导致模型训练效果不佳;而过大的 num_unit 会带来过拟合并且训练时间过长的缺点。由图 2 可知,当词向量维度为 512 时模型效果最优,因此对固定词向量维度取为 512,其他参数为默认参数,取 num_unit 为 32,64,128,256,512 进行模型训练尝试找到最佳的取值。基于 ppl 及 bleu 进行对照的折线变化如图 3 所示。

由图 3 可见,在训练初期,num_unit 高的模型与 num_unit 低的模型的 ppl 变化折线图像会有一个交点。在交点以前,num_unit 越小,ppl 越小。在交点以后,num_unit 越小,ppl 越大。当训练步数足够大时,随着 num_unit 的增大,最终的 ppl 变小。num_unit 越大,bleu 变化折线图整体在 num_unit 小的 bleu 变化折线图上方。



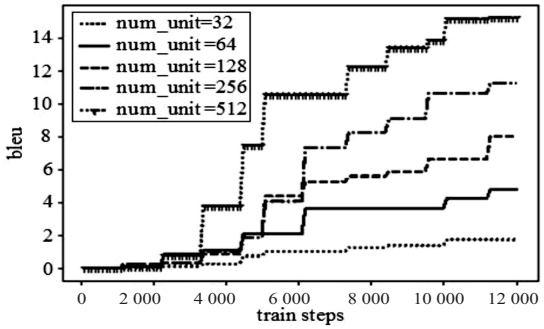


图 3 不同 num_unit 模型的 ppl 及 blue 对照折线变化

综上所述,在 num_unit 取值为 32,64,128,256,512 时,num_unit 为 512 时的模型训练效果最优。并根据图像显示,继续增大 num_unit 的取值,模型有进一步优化的可能。

5.4 超参数 batch_size 对模型训练的影响

超参数 batch_size 代表了一次输入给神经网络的样本数,在合理范围内,越大的 batch_size 使得参数修正的方向越准确,震荡越小;而过大的 batch_size 会使得一次 epoch 内所需的迭代次数变少,从而对参数的修正变得更加缓慢;而过小的 batch_size 会使得随机性较大,震荡较大,难以达到收敛。将固定词向量维度设为 512,num_unit 为 512,其他为默认参数,由于机器性能的限制,无法选用更大的 batch_size 进行实验,因此分别选取 batch_size 为 4,32,64,128,192,256 进行训练尝试找到 batch_size 的合理范围和最佳的选值。

基于 ppl 及 bleu 进行对照的折线变化如图 4 所示。

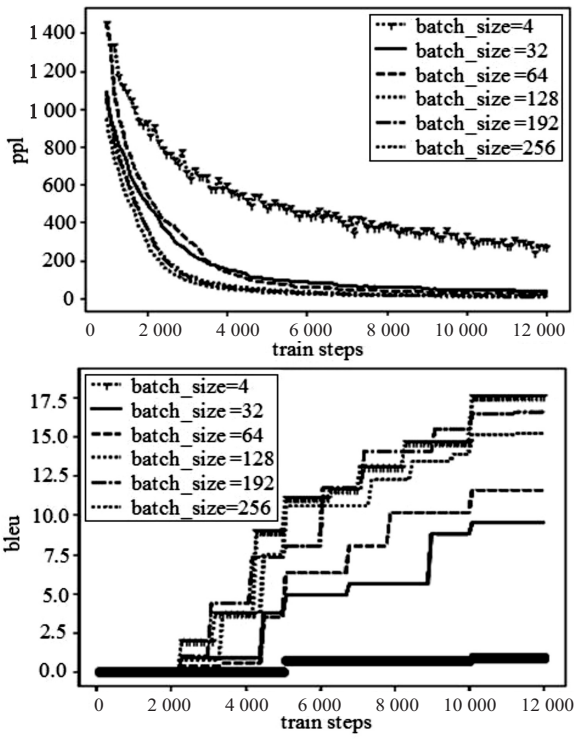


图 4 不同 batch_size 模型的 ppl 及 bleu 对照折线

由图 4 可见,当 batch_size 过小时,曲线震荡较大,模型的训练效果较差。当 batch_size 在范围[32,256]内时,曲线震荡较小,模型训练效果相近,因此范围为[32,256]是 batch_size 的合适选值范围。当 batch_size 过小时,模型训练效果非常差。随着 batch_size 的增大,最终的 bleu 值变大。

综上所述,batch_size 在取值为 4,32,64,128,192,256 时,batch_size 为 256 时的模型训练效果最优。并根据图像显示,继续增大词向量维度,模型有进一步优化的可能。

5.5 最终的模型训练

根据以上的实验结果,最终选取词向量维度为 512,num_unit 为 512,batch_size 为 256 进行最终的模型训练。

模型的 bleu 值详细变化如图 5 所示。

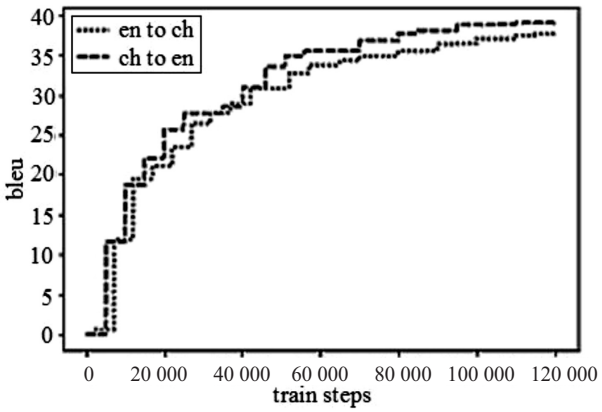


图 5 模型的 bleu 详细变化折线

由图 5 所示,当模型训练步数到达 120 000 步左右时,bleu 渐渐达到峰值。

6 翻译结果

将该文的英译汉模型与百度翻译进行翻译效果对比,如表 3 所示。

由表 3 可以看到,使用训练模型翻译的语句均能将专业词汇翻译正确,而百度翻译会将专业词汇翻译错。如测试语句 1 中的“binary search”翻译成了二进制搜索,测试语句 2 中的“greedy”翻译成了贪婪,测试语句 3 中的“memoization”更是直接没有翻译。由此可见使用文中训练的翻译模型在计算机算法类领域比百度翻译的效果更好。

7 结束语

该文主要进行了计算机算法类垂直领域的中英翻译优化。以关键词进行检索并爬取中英双语例句,通过收集与计算机算法类相关度高的文本优化翻译效果。对英文文本进行小写化处理,避免词汇表出现重

复单词,降低训练成本以及优化训练效果。通过 jieba 库对中文文本分词和 nltk 库对英文文本分词。通过制作计算机算法类分词字典,增加计算机算法类文本的分词正确率。通过 Word2Vec 算法训练不同维度的词向量,进行词汇间文本相似度测试得出,词向量维度越大,词汇间相似度越低,但所得出的测试结果越具有实际意义。通过实验分析发现,词向量维度为 512,num_unit 为 512,batch_size 为 256 时模型效果最佳。

表 3　翻译对比结果

测试语句	文中模型翻译结果	百度翻译结果
tests proved this method is simple and more effective than binary search method .	通过测试,该方法简单且比二分搜索方法效率还高	实验证明,该方法比二进制搜索方法简单、有效
a mathematical model is firstly proposed . then the greedy heuristic algorithm is applied to solve the model .	首先建立此问题的数学模型,并采用启发式贪心算法对模型进行求解	首先提出了一个数学模型。然后采用贪婪启发式算法对模型进行求解
the solution of packrat parsing is to use memoization , i. e. storing intermediate parsing results , instead of calculating these results over and over .	Packrat 解析的解决方法是使用记忆化,例如将解析的中间结果保存下来,而不是一遍一遍的重新计算	packrat 解析的解决方案是使用 memoization,即存储中间解析结果,而不是一遍又一遍地计算这些结果

参考文献:

[1] 刘　洋. 神经机器翻译前沿进展[J]. 计算机研究与发展,2017,54(6):1144-1149.

[2] LI H. Deep learning for natural language processing:advantages and challenges[J]. National Science Review,2017,5(1):24-26.

[3] CHO K,VAN MERRIËNBOER B,GULCEHRE C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv:1406.1078,2014.

[4] 肖新凤,李石君,余　伟,等. 基于改进 seq2seq 模型的英汉翻译研究[J]. 计算机工程与科学,2019,41(7):1257-1265.

[5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th international conference on neural information processing system (NIPS 2014). Cambridge, MA, USA: MIT Press,2014:3104-3112.

[6] 董陆森. 机器翻译中的常用神经网络模型[J]. 电子技术与软件工程,2018(10):147.

[7] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473,2014.

[8] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in neural information processing systems 30:annual conference on neural information processing systems 2017. Long Beach,CA,USA:[s. n.],2017:5998-6008.

[9] WU Y,SCHUSTER M,CHEN Z,et al. Google's neural machine translation system:bridging the gap between human and machine translation[J]. arXiv:1609.08144,2016.

[10] 刘　群,俞士汶. 汉英机器翻译的难点分析[C]//1998 中文信息处理国际会议. 北京:出版者不详,1998:507-514.

[11] 于重重,操　镭,尹蔚彬,等. 吕苏语口语标注语料的自动分词方法研究[J]. 计算机应用研究,2017,34(5):1325-1328.

[12] HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. Amherst:Lawrence Erlbaum, Hillsdale,1986:1-12.

[13] RONG X. word2vec parameter learning explained[J]. arXiv:1411.2738,2014.

[14] 周　练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济,2015,25(2):145-148.

[15] PAPINENI K,ROUKOS S,WARD T,et al. BLEU:a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for computational linguistics. Pennsylvania, USA: ACL,2002:311-318.