文章编号：2095-2163(2020)07-0044-05 中图分类号：TP399 文献标志码：A

跨语言语义向量的生成模型

金卓林，朱聪慧
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150000）

摘 要：目前较优秀的 NLP 系统模型比较依赖有标注的数据来学习复杂的模型，这种模型通常在一个单一语料上进行训练，不能直接利用到其他语言上。收集每种语料上的训练数据是不现实的，因此想通过跨语言的方式进行低资源语料之间的迁移学习，达到在无监督学习的条件下能够进行跨语言的任务，这里进行了句子级别的语义向量的生成，并利用下游分类任务查看语义向量的质量。基于此本文提出了基于跨语言语义向量生成的模型，并引入命名实体识别，利用平行语料做语义对齐等多任务学习。实验数据为 XNLI 数据集，也是跨语言任务中常用的数据集。在多任务学习模型下，和基线模型相比，在 XNLI 数据集上效果有明显提升。

关键词：跨语言任务；迁移学习；多任务学习；语义向量

Generation Model of Cross-lingual Sentence Embedding

JIN Zhuolin, ZHU Conghui
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China)

【Abstract】 Nowadays excellent NLP system model relies on labeled data to learn complex models, which are usually trained on a single corpus, but cannot be directly used in other languages. Because it is unrealistic to collect the training data on each kind of corpus, we want to carry out the transfer learning between low resource corpus in a cross-lingual way, so that we can carry out the cross-lingual task under the condition of unsupervised learning. Here we generate sentence embedding, and use the downstream classification task to check the quality of it. Based on this, this paper proposes a cross-lingual sentence embedding generation model, and introduces named entity recognition and semantic alignment using parallel corpus. The experimental data is XNLI data set, which is also commonly used in cross language tasks. In the multi task learning model, compared with the baseline model, the effect on the XNLI data set is significantly improved.

【Key words】 cross-lingual; transfer learning; multi-task learning; sentence embedding

0 引 言

目前较优秀的 NLP 系统模型比较依赖有标注的数据来学习复杂的模型，这种模型通常在一个单一语料上进行训练，不能直接利用到其他语言上。收集每种语料上的训练数据难度很大，因此希望借助于语料充足的其他语言数据来做没有具体训练数据的语言上的任务，也就是进行无监督的目标语言学习。由于自然语言任务的高度理解，选取的任务自然语言理解作为迁移学习的任务。其他主流的方法构建在词级别的对齐后，利用对齐后的词向量来做句子向量的生成，会有偏差。

本文将比较不同结构对句子进行编码的效果，经过试验在跨语言自然语言理解任务上采用性能最好的模型作为基线模型。并利用平行语料来对齐源语言编码器和目标语言编码器，实现无监督学习，再将编码器的性能通过多任务学习的方式进一步提高，最终得到的跨语言无监督学习得到的语义向量的性能接近有监督学习得到的源语言的语义向量。在公开实验数据集 XNLI 上，多任务跨语言词向量生成的方法比其他模型有明显的提高。

1 相关工作

生成跨语言的语义向量中很多的工作都是建立在词级别的，有一些方法是去学习跨语言词向量，将目标语言的词向量映射到源语言的词向量空间上。大部分词级别的方法也都需要有监督的数据，来对齐不同的词向量。最近也有无监督的方法，Mikolov 在 13 年提出了向量空间可以编码词汇间的关系，并提出不同语言间的集合关系是类似的[1]。Mikolov 利用了 5000 对最常用的反义词对作为监督来学习，式(1)。

$$\min \sum_{i=1}^{n} |Wx_i - z_i|^2. \quad (1)$$

有很多方法把词级别的表示扩展到句子以及文档级别的表示，其中最直接的方法就是利用所有词

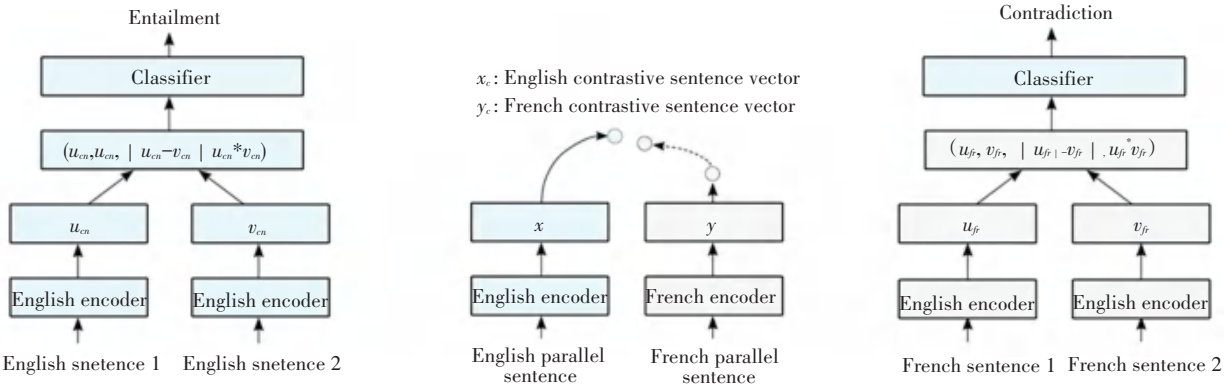
作者简介：金卓林(1995-)，男，硕士研究生，主要研究方向：自然语言处理、机器学习；朱聪慧(1981-)，男，博士，副教授，硕士生导师，主要研究方向：自然语言处理、机器翻译。

收稿日期：2020-05-06

向量的平均值或加权平均值，被称为连续词袋模型。尽管这种方法看起来很简单，但通常却能作为一个很强的基线模型，也可以利用无监督的训练 word2vec 中扩展 skip-gram 的方式利用词向量；也有一些在预训练的语言模型上利用这些固定大小的句子级别的向量表示，利用这些模型的每个隐状态作为上下文环境的词向量，来对其他任务进行微调。Chandar 等人训练了双语的自动编码器，其中损失函数是为了减小两个语言之间重构的误差[2]；Schwenk 等人在序列到序列的机器翻译模型中联合一起训练多种语言，来学习一个共享多语言的句子级别的向量空间表示[3]；Hermann 等人提出了利用单个单词和两个单词的组合来学习文档级别的表示[4]。

2　跨语言语义向量的生成

本文提出的模型以循环神经网络作为编码器，利用平行语料来对齐源语言端的编码器和目标语言端的编码器，对齐后的目标语言端编码器可以直接在目标语言端测试。另外，利用多任务学习进一步提高了编码器的性能，本文提出的模型如图 1 所示。



(a) Learning NLI English encoder and classifier　(b) Aligning sentence encoders with parallel data　(c) Inference in the other language

图 1　跨语言句子级别向量生成的模型

Fig. 1　Model of cross-lingual sentence embedding

2.1　问题定义

通过语义向量的编码器将整个句子编码成固定长度的向量，且得到的语义向量能够用于下游任务。形式化表示模型的输入表示为：

$$x_i = (x_1, x_2, \cdots, x_n)$$

其中，x_i 为输入的第 i 个词，n 为句子中所有词的数量。经过编码器处理后，得到固定长度为 d 的向量，用来表示句子的语义信息。

2.2　编码器模型

本文使用 BiLSTM + maxpooling 作为计算得到句子级别语义向量的编码器。LSTM 的全称是长短时记忆神经网络，适合对时序序列建模。LSTM 模型是由 t 时刻的输入词 x_t，细胞状态 C_t，隐层状态 h_t，遗忘门 f_t、记忆门 i_t、输出门 o_t 组成。

计算过程是利用遗忘门和记忆门对输入信息中有效信息进行保留，无效信息则会被移除。记忆门的输入是上一时刻的隐层状态 h_{t-1}、当前时刻的输入信息 x_t。其中临时细胞状态 $\hat{C}_t$ 的输入和记忆门相同，用公式表达为式(2)和式(3)：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{2}$$

$$\hat{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{3}$$

利用遗忘门 f_t 选择对上一时刻的细胞状态进行保留，利用记忆门对临时细胞状态进行保留，用公式(4)表示：

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t. \tag{4}$$

计算输出门 o_t 是利用上一时刻的隐层状态 h_{t-1} 和 t 时刻输入的词 x_t。利用输出门 o_t 和当前细胞状态 C_t 可以计算得到当前时刻的隐层状态，用公式(5)和公式(6)表达：

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o), \tag{5}$$

$$h_t = o_t * tanh(C_t). \tag{6}$$

将前向和后向的 LSTM 拼接在一起就得到了双向的 LSTM(BiLSTM)，公式得到每个时刻 t 的隐藏状态 h_t 后，经过一个 max pooling 得到最终的句子表示。其中，max pooling 的意思是将每个时刻 t 下 h_t 对应维度上的值进行比较，max 表示取对应维度上最大的值。

除了 BiLSTM + max pooling 以外，还比较了 GRU、CNN 和 Innerattention 模型在 SNLI 数据集上的效果，结果见表 1。

表 1 SNLI 数据集实验结果

Tab. 1 Experiment of SNLI

方法	R-1
BiLSTM + Mean pooling	81.9
CNN	83.0
Inner Attention	77.9
BiLSTM+Max pooling	**84.2**

其中 SNLI 是建立在英文语料上的自然语言理解常用的数据集,从实验结果可以看到 BiLSTM + maxpooling 的方法效果最好,因此选择使用这种模型作为本文模型的编码器。

2.3 跨语言句子级别语义向量的生成方法

对于跨语言任务最直接的方法就是可以利用机器翻译的方式,将数据翻译后,再用来训练。这种方法都能够作为很强的基线模型,但都有自己的缺点:第一,重新训练新的模型;第二,比较依赖测试阶段的语义计算。另外这种方法需要比较好性能的机器翻译系统,机器翻译模型的性能影响着翻译后语料的质量。

本文提出了一种基于自然语言理解的跨语言句子向量生成的方法:首先,在有充足训练数据的语料上,训练 NLI 的模型,作为指导模型,比如在有充足训练数据的英语上,训练出一个准确率达到 90%以上的模型,采用的是 BiLSTM 对每一个时间步取最大池化操作,将得到的结果作为整个句子的语义向量。和目前其他最好的模型相比,这个轻量级的模型得到的结果已经接近最好的模型,将这个模型保存下来,分为编码层和分类层,其中编码层是将整个句子编码成一个固定大小的语义向量,分类层是将语义向量经过全连接,从语义向量的维度转变到分类数目大小的维度。将这个模型保存下来,并且将所有的参数都固定住,相当于只用这个模型进行预测;其次,利用第一个阶段得到的固定参数的源语言端的编码器,和少量平行语料进行训练,得到目标语言端的编码器,使用和源语言端相同结构的编码器进行拟合。这里将一句平行语料输入到两个编码器中,分别得到两种语料上的语义向量,计算两个向量之间的距离,使用的损失函数(7):

$$sim(x,y) - \lambda(sim(x_c,y) + sim(x,y_c)). \quad (7)$$

其中,sim 是用来计算两个向量间距离的函数,包括 L1 距离、L2 距离、KL 散度、余弦线相似度; x_c 和 y_c 是随机选取的负样本, x 和 y 是一对平行语料,这里的损失函数的意义是对 x 和 y 的语义向量距离去 L2 距离, 让 x 和 y 在空间上很接近,由于产生 x 的编码器的所有参数都被固定了,因此产生 y 的编码器经过反向传播后会向 x 的空间拟合,由于 x_c 和 y_c 是随机选取的负样本, x_c 和 y 不是一对平行语料,同样 x 和 y_c 也不是一对平行语料,因此计算得到的距离应该尽可能的大, 产生目标语言的编码器会尽可能让 $y(y_c)$ 离 $x_c(x)$ 更远。 由于负样本选取的随机性,在选取负样本时,本文进行了多次选取负样本,利用多次选取后的平均值作为通用的负样本。利用大量平行语料进行对齐,从实验结果来看,使用平行语料的数量越多,结果会越好。将得到的目标语言端的编码器的模型参数保存下来,用于第三个阶段。

在第三个阶段中,由于两个编码器经过对齐的操作后,理论上得到的向量结果应该高度一致,因此可以将目标语言的编码器替代源语言端的编码器,将目标语言的编码器和在第一阶段得到的分类层拼接,直接预测结果。经过这三个阶段,只利用平行语料就可以进行没有目标语言数据的训练。

2.4 多任务下跨语言语义向量的生成

生成后的跨语义语义向量经过在目标语言端的测试,和源语言端的准确率还有一定的差距,在这个任务下是有可以提高的空间的。因此,除了使用 NLI 作为指导模型以外,还考虑使用命名实体识别作为训练数据。采用多任务学习的方法,更加全面的利用到其他易获取的数据资源。本文利用了一个多任务的双向编码器框架进行训练,不仅能够保持良好的迁移学习的能力,也能很好地完成语义信息的转化,可以直接加载到其他任务上。任务包括:自然语言理解任务、命名实体识别任务以及进行迁移学习的翻译任务。其中自然语言理解任务依然使用原结构,利用 BiLSTM 对句子双向编码,取每个时间步的最大池化结果作为最终结果。命名实体识别任务的模型如图 2 所示。

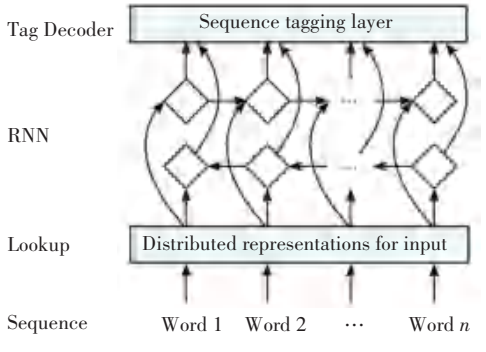


图 2 命名实体识别任务模型图

Fig. 2 Model of NER task

在多任务训练中,NLI 的预测作为任务一,命名

实体识别任务作为任务二，利用平行语料对齐句子作为任务三，三个任务共用同一个编码器。分别将数据传入到共享参数的编码器中，分别计算各自任务上的损失函数结果，将 loss 分别做反向传播。更新编码器的参数，最终将得到的编码器参数固定下来，加上源语言端的分类层，直接进行在法语上的预测结果。其中多任务下的流程如图 3 所示。

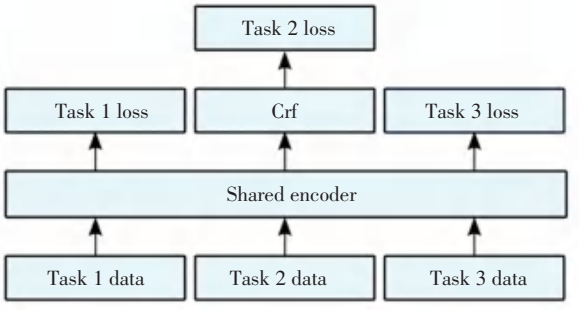


图 3　多任务学习的流程

Fig. 3　Flow chart of multi-task learning

3　实验设计与分析

3.1　数据集

数据集选取 XNLI 数据集，即跨语言任务上的 NLI 的数据集。XNLI 一共将 NLI 数据集扩展到 15 种语言，并以 NLI 的三分类格式为每种语言分别提供了 7 500 个经人工标注的开发和测试实例，合计 112 500 个标准句子对。XNLI 中包括的语言跨越了多个语系，特别还包括了斯瓦希里语和乌尔都语这两种语料资源稀缺的语言。

另外，利用 WMT14 英语到法语的 360 万对平行语料作为数据集进行对齐编码器。

在多任务实验下，使用 DAWT 数据集作为命名实体识别任务的数据集，是一个有多种英语的命名实体识别数据集，包括英语、西班牙语、意大利语、法语和德语等。一共 175 万句数据，其中选取 80%作为训练集，10%作为验证集，10%作为测试集。

3.2　实验参数设置

在实验中使用 Pytorch 框架作为本实验框架，采取 Adam 作为模型训练的优化器，其中设置参数 adam-betas 为(0.9, 0.98)，预热训练 4 000 轮迭代，dropout 采取 0.3，权重衰减为 0.000 1，损失函数采用类交叉熵，做标签顺滑，参数为 0.1。batch 大小为 256，在训练过程中加入 early_stop 机制。

3.3　实验结果及分析

本文用在目标语言法语上的分类准确率作为评价指标。分别计算负样例选取为 1、5 和 10 的情况下，模型在 XNLI 数据集上的实验结果，见表 2。

表 2　XNLI 数据集实验结果

Tab. 2　Result of XNLI

Sim	K = 1	K = 5	K = 10
Cos	33.3	20.0	27.0
Cos&KL	**57.4**	**64.32**	**65.14**

通过表 2 可以看到，由于随机选取负样本的随机性过高，当多次选取负样本后，选取平均值的方法可以使得结果有提升，且选取负样例数目越多，结果越好。

在负样例选取 K = 10 时，比较不同计算向量距离方法的不同，对实验结果的影响，在 XNLI 数据集上的实验结果见表 3。

表 3　XNLI 数据集实验结果

Tab. 3　Result of XNLI

Result of XNLI 方法	Acc
L2	50.57
KL	58.65
Cos	63.29
Cos+KL	65.12
Cos+KL+L1+L2	**67.58**

通过表 3 可以看到，计算向量间距离的单一会造成结果较差，向量结合方式越多，计算得到的语义向量效果更好。

结合分类、序列标注以及机器翻译共同训练，其中所有句子共享句子级别的编码层，在一轮训练中分别计算三个任务的损失函数并反向传播。分别比较了 5 轮训练后在命名实体识别任务上的 F 值、在英语端的准确率和在法语上的准确率。实验结果见表 4。

表 4　多任务学习下实验结果

Tab. 4　Result of Multi-task learning

Epoch	F	En	Fr
1	46.21	87.65	**77.75**
2	50.05	91.21	75.76
3	51.26	91.43	76.92
4	51.12	91.54	76.12
5	**52.03**	**91.6**	76.92

可以看到随着训练轮数的增多，在命名实体识别和英语上的准确率都有所升高，但是在法语端的准确率在下降。和上小节相比，结果有了比较大的提升，甚至接近了在源语言端的准确率。在多任务学习的设置下，在目标语言端的准确率有明显的增长。

4 结束语

本文比较了不同结构对句子编码的效果,采用最好的结构在跨语言任务上的实验作为基线模型,再将编码器的性能通过多任务学习的方式进一步提高,最终跨语言无监督学习得到的语义向量的性能接近有监督学习得到的源语言的语义向量,实现了在没有目标语言数据的情况下,无监督的学习,且结果和基线模型相比有很大的提升。这对于一些稀缺数据的语种有极大的意义,意味着不需要额外的人力物力去标注稀缺语种的数据。

参考文献

[1] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in neural information processing systems. 2013: 3111-3119.

[2] RONGALI S, SARATH CHANDAR A P, RAVINDRAN B. From multiple views to single view: a neural network approach[C]// Proceedings of the Second ACM IKDD Conference on Data Sciences. 2015: 104-109.

[3] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv preprint arXiv: 1509.00685, 2015.

[4] GULCEHRE C, FIRAT O, XU K, et al. On using monolingual corpora in neural machine translation[J]. arXiv preprint arXiv: 1503.03535, 2015.

[5] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. arXiv preprint arXiv:1602.06023, 2016.

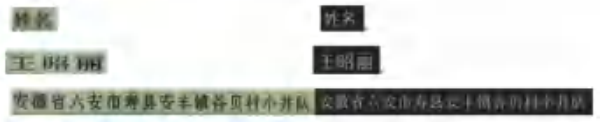
[6] KAYACAN, E, KAYACAN, E, RAMON, H. Learning in Centralized Nonlinear Model Predictive Control: Application to an Autonomous Tractor - Trailer System [J]. Control Systems Technology, IEEE Transactions on, 2015, 23(1):197-205.

[7] GULCEHRE C, AHN S, NALLAPATI R, et al. Pointing the unknown words[J]. arXiv preprint arXiv:1603.08148, 2016.

(上接第 43 页)

模型,读入待预测的图像,resize 高为 32 的灰度图像,将该图像送入网络,再将网络输出解码成文字即可输出。

使用 CRNN 进行文字识别,识别结果如图 6 所示。



(a) 文字区域 (b) crnn 识别结果

(a) Text area (b) Recognition result of crnn

图 6 CRNN 识别结果示例

Fig. 6 Recognition result example of CRNN

由图 6 可见,对于驾驶证文字识别,用 CRNN 可以准确识别出证照中文字。

为了验证本文文字识别算法的有效性,实验使用 280 张手机拍摄的驾驶证图片,用于计算关键区域的识别准确率。统计正确识别出的文字个数、错误检测到的文字个数以及未检测到的文字个数,通过和实际输入的文字个数的比值得到算法的准确率。统计结果见表 2。

表 2 驾驶证识别准确率

Tab. 2 Driver License Recognition Accuracy

驾驶证文字识别	识别准确率/%
姓名	95.7
证号	94.3
地址	94.8

实验结果表明,基于深度学习的驾驶证识别算法在驾驶证文字检测与识别都具有良好的效果。

3 结束语

为了对驾驶证图片进行文字识别,本文提出了一种基于深度学习的机动车驾驶证的检测与识别算法。针对驾驶证图片存在的背景复杂、角度倾斜的问题,利用一系列图像预处理方式对原始图像进行校正;使用 CTPN 算法对校正结果进行文字检测;使用 CRNN 算法对检测出的文字框进行识别。本文对驾驶证文字检测和识别均采用深度学习方法实现,解决了传统证件识别需要进行版面分析,制作模板的问题。

实验结果证明算法在识别准确率和鲁棒性等方面具有优势,达到实用性的标准,今后将针对更为复杂的拍摄场景进行进一步优化。

参考文献

[1] 林涵阳,詹永照,陈羽中. 复杂场景中机动车行驶证快速检测与识别[J]. 小型微型计算机系统,2019,40(5): 1076-1082.

[2] 白翔,杨明锟,石葆光,等. 基于深度学习的场景文字检测与识别[J]. 中国科学(信息科学),2018,48(5): 531-544.

[3] 段金宝. 基于深度神经网络的证件图像文本识别方法[D]. 北京:北京邮电大学,2018.

[4] 李亮. 基于 Tesseract_OCR 的驾驶证识别系统设计与实现[D]. 成都:电子科技大学,2018.

[5] 蒋冲宇,鲁统伟,闵峰,等. 基于神经网络的发票文字检测与识别方法[J]. 武汉工程大学学报,2019,41(6): 586-590.

[6] 张翩. 复杂背景下证件识别技术的研究与实现[D]. 电子科技大学,2017.