语言先验知识对神经自然语言处理任务的影响*

贝超[1,2]，胡珀[2]

（1. 中译语通科技(北京)有限公司，北京市 100043；2. 华中师范大学，湖北省 武汉市 430079）

摘要：随着互联网的发展及硬件的更新，神经网络模型被广泛应用于自然语言处理、图像识别等领域。目前，结合传统自然语言处理方法和神经网络模型正日益成为研究的热点。引入先验知识代表了传统方法的惯例，然而它们对基于神经网络模型的自然语言处理任务的影响尚不清楚。鉴于此，本文尝试探究语言层先验知识对基于神经网络模型的若干自然语言处理任务的影响。根据不同任务的特点，比较了不同先验知识和不同输入位置对不同神经网络模型的影响。通过大量的对比实验发现：先验知识并不是对所有任务都适用，在神经网络模型的合适位置加入合适的先验知识方可加快模型的收敛速度，提高相关任务的效果。

关键词：神经网络；自然语言处理；先验知识

Influence of Prior Knowledge on Neural Networks Model in NLP Tasks

Abstract: With the development of the Internet and the updating of computer hardware, the neural network model is widely used in natural language processing, image recognition and other fields. At present, the combination of traditional NLP methods and neural network model is becoming a hot research trend. It is still unclear whether the prior knowledge which is the practice of traditional methods has impact on the task of neural network model. In this paper, we have explored the influence of linguistic prior knowledge on neural network models in several NLP tasks. According to the characteristics of different tasks, we compared the effects of the different prior knowledge and the different input location on the different neural network models. Through a large number of comparative experiments, the results show that in some reasonable locations of some neural network, the prior knowledge can speed up the model's convergence speed and improve the result, while it is not applicable for all conditions.

Key words: neural network；natural language processing；prior knowledge

1 引言

互联网的飞速发展逐渐改变了普通大众的生活，各类信息呈指数级速度增长，为研究者们提供了更丰富的实验数据，同时也为神经网络模型的训练提供了充足的样本。

如今神经网络已深入到图像识别、语音识别、自然语言处理等众多领域，结合大规模的训练集以及高性能的图形处理器，使得在很多领域相较于传统方法取得了较大的突破。2012年 Hinton 小组第一次参加 ImageNet 比赛便以领先第二名 10%以上的成绩夺冠，使用的正是深层卷积神经网络，其系统正是著名的 AlexNet[1]。2016 年，谷歌发布了神经网络机器翻译系统，在部分语言对上已接近人工翻译的质量。在部分领域，神经网络业已成为目前的主流甚至最佳方法。

然而在自然语言处理领域，神经网络所需训练集数据量大和训练速度慢的问题并未得到根本性解决，使得它们在与传统自然语言处理方法进行比较时的优势并没有如此明显。因此，如何将传统方法与神经网络结合起来，优势互补，成为了如今神经网络研究的一大趋势。虽然已有工作在神经网络模型中加入了部分先验知识，然而它们并没有专门探究先验知识对神经网络模型的具体影响。

本研究考察结合传统方法所惯用的先验知识，探究语言层先验知识对若干自然语言处理

*基金项目：国家自然科学基金青年基金项目（61402191）；华中师范大学中央高校基本科研业务费教育科学专项资助项目（CCNU16JYKX15）；国家语委科研项目（WT135-11）

作者简介：贝超（1992-），男，硕士，主要研究领域为自然语言处理；胡珀（1980-），男，博士，副教授，通讯作者，主要研究方向为自然语言处理、自动文摘。

任务中神经网络模型的影响。我们根据不同的任务特点，探究了不同先验知识和不同先验知识的输入位置对神经网络模型的影响，以及先验知识对不同神经网络模型的影响。

2 相关工作

2.1 神经网络在自然语言处理基本任务中的应用

自然语言处理的基本任务主要包括词性标注、chunk 标注、命名实体识别、语义角色标注等。对于词性标注问题，Toutanova 等[2]使用最大熵分类器并在标注时使用双向依存网络，其准确率达到 97.24%。对于 chunk 标注，Shen 和 Sarkar[3]使用投票分类器方案，其中每个分类器在不同的标签上进行训练，其结果的 F1 分数达到 95.23%。对于命名实体识别，Ando 和 Zhang[4]使用了非监督的方法，其结果的 F1 分数达到 89.31%。对于语义角色标注，Koomen 等 [5]使用 Winnow-like 分类器结果的 F1 分数达到 77.92%。以上工作均为使用传统机器学习的方法来处理 NLP 基本任务，可以注意到传统方法使用的特征大部分互为先验知识，这样使用额外的特征在传统机器学习方法中很常见。

直到 2011 年，Collobert 等[6]使用神经网络的方法做自然语言处理基本任务，使其准确率基本达到传统方法的准确率。在无需任何额外的自然语言处理知识的情况下，他们使用可应用于四种不同任务的模型，其结果已接近传统方法的准确率，其主要贡献在于降低了自然语言处理的门槛，同时多个任务只需一个模型结构。并且，他们还在神经网络中加入了词特征，如词的大小写，词缀信息等，这些词特征就是额外加入的先验知识。2013 年 Zheng 等使用神经网络模型和维特比解码进行中文分词和词性标注[7]，同时提出了一种类似感知器的训练神经网络的方法来加速训练。虽然工作中并未加入更多的先验知识，但是在结论中已提到未来可以加入更多先验知识。

2.2 神经网络在文本分类任务中的应用

文本分类是自然语言处理的经典任务之一，主要使用计算机对文本集按照一定的分类体系或标准进行自动分类标记。传统机器学习方法处理文本分类问题主要分为文本表示和机器学习分类方法。

对于文本表示，主要使用词袋模型。词袋模型以词为单位来表示文本，是传统机器学习方法进行文本分类的标准模式。而以词为单位，以词频为基础计算权重，如 TFIDF[8]，来表示文本也是常见做法。TFIDF 是一种统计的方法，即一个词在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为该词具有很好的区分能力，以此作为该词的权重，可以很好地用作分类。

对于机器学习分类方法，支持向量机[9]是二分类最有效的方法。虽然支持向量机需要大量的存储资源和高的计算能力，但是其分隔模式可以有效地克服样本分布、冗余特征以及过拟合等问题的影响，具有很好的泛化能力。支持向量机主要通过非线性映射，把样本空间映射到一个高维甚至无限维的特征空间中，使得在原本的样本空间非线性可分的问题转化为在高维的特征空间的线性可分问题。

除了传统的机器学习方法，Siwei Lai 等[10]提出循环神经网络和卷积神经网络结合的方法进行本文分类。其模型使用双向长短时记忆循环神经网络得到信息与词向量来构建表示文本，之后进行卷积神经网络的构建。此外，借鉴图像识别问题的处理方法，为了提取文本中更深层次的信息，Zhang 等[11]使用字符级的卷积神经网络进行文本分类，以字符为基本单元取得了接近以词为基本单元方法的效果。其以 70 个字符为基础，包括 26 个英文字母，10 个数字以及其他字符来表示文本，使用多层卷积神经网络和全连接神经网络来挖掘文本的信息。

虽然上述应用神经网络模型的文本分类中，并不需要人工选择特征，可以自动学习特征，

并且可以使用类似的神经网络模型处理大部分的文本分类应用。然而以上工作中均未加入先验知识。而考虑语言层先验知识，在此基础之上进行更深层次的提取特征将可能有利于模型获得更有用的信息。

2.3 神经网络在机器翻译任务中的应用

在神经网络机器翻译出现以前，一直都是基于短语的机器翻译作为主流。Philipp Koehn 等[12]使用基于短语的统计机器翻译得到了工业界的广泛应用，但其缺点比较明显。虽然基于短语的统计机器翻译相对确定了部分语序，然而从结果来看其翻译结果的语序并没有达到目标语言语序的基本要求，尤其在语序的差异较大的语言对中，其翻译结果很难让人理解。同时训练过程复杂，对于机器翻译知识的要求比较高。

神经网络模型同样使用统计的方法，但模型结构为神经网络。Sutskever 等[13]提出使用基于长短时记忆循环神经网络的序列到序列模型，可以应用于机器翻译、文本摘要等序列生成问题。Bahdanau 等[14]在 Sutskever 等人工作的基础上加入注意力机制，提出基于注意力机制的编码到解码模型，其结构尤其适合机器翻译任务。

然而，如谷歌神经机器翻译系统这样成功的工业界应用也并未加入更多的先验知识，但对于机器翻译任务来说，双语中可供挖掘的信息相比于其他任务往往更多，毕竟双语的先验知识明显较单语更丰富，这使得机器翻译任务具有更多的可供挖掘的信息，具有相比其他部分任务更大的提升空间。而加入语言层先验知识，可以抓住源语言和目标语言之间的某些联系，提高源语言与目标语言词与词之间的紧密性以获取更多的信息，来帮助翻译系统提高翻译质量。

3 语言层先验知识对若干自然语言处理任务中神经网络模型的效果影响

为了探究不同先验知识和不同先验知识输入位置对神经网络模型的影响以及先验知识对不同 NLP 神经网络模型的影响，我们设计了三类难易程度不同的任务：即自然语言处理基本任务、文本分类和机器翻译，并根据不同的任务特点分别探究了以上三个动因。

3.1 先验知识对 NLP 基本任务中神经网络模型的影响

3.1.1 实验数据

考虑到自然语言处理基本任务的特点，基本任务的实验数据选用布朗语料库。我们设置其中 450 个文本作为训练集，开发集和测试集各为 25 个文本。需要注意的是，语料中已标注了词性信息，命名实体信息则需要使用传统自动标注工具来标注。

3.1.2 数据预处理

使用 Mosesdecoder 进行分词，并通过斯坦福自然语言处理工具对其进行命名实体识别，与词性互为额外的特征输入和标注类别。实验主要采用窗口模型结构，通过周围词给中间词标注。

3.1.3 神经网络模型设计及参数设定

我们主要采用全连接神经网络模型和基于长短时记忆的循环神经网络模型，采用类似 Collobert 等人提出的窗口方法，其具体参数设定如下：

全连接神经网络结构如图 1 所示。首先以词向量作为输入，窗口大小为 5，经过 5 层全连接神经网络挖掘其特征，此后通过 softmax 层做标注。其中词向量大小为 128，词表大小为 6000，全连接神经网络隐藏层大小为 128，且在全连接神经网络层中加入 dropout 层，其随机断开比为 50%。额外的特征向量接在词向量的后面，根据特征类别数量，词性向量大

小为 40，命名实体向量大小为 4。

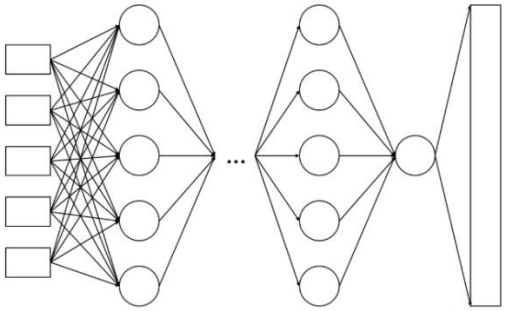


图 1 全连接神经网络的标注模型

循环神经网络结构如图 2 所示。首先以词向量作为输入，窗口大小为 5，经过 1 层循环神经网络挖掘其特征，此后通过 2 层全连接神经网络做标注。其中词向量大小为 128，词表大小同样为 6000，循环神经网络使用的是长短时记忆网络，循环神经网络和全连接神经网络的隐藏层大小均为 128，在两层全连接神经网络中间加入 dropout 层，其随机断开比为 30%。同样的，额外的特征向量紧接在词向量后面，词性向量大小为 40，命名实体向量大小为 4。

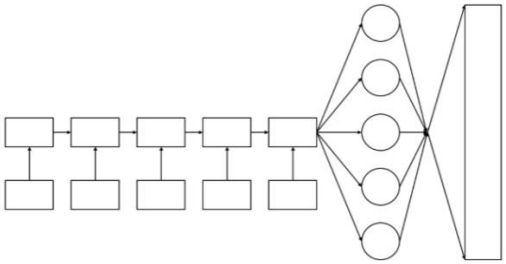
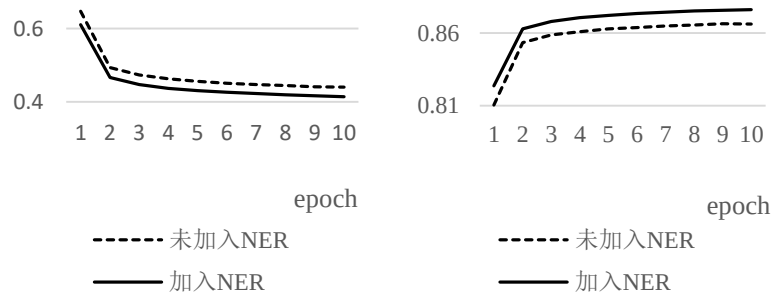


图 2 循环神经网络的标注模型

3.1.4 实验结果与分析

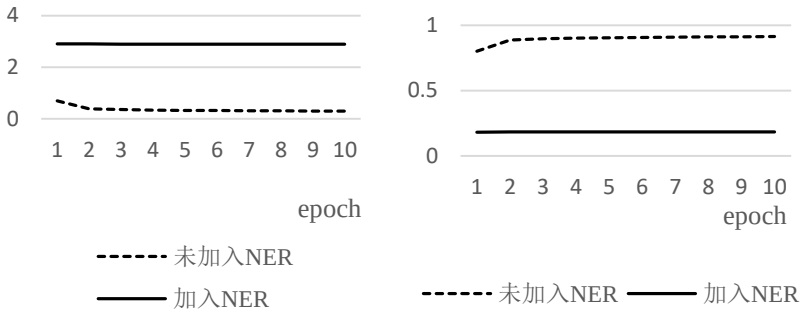
我们分别进行了四组对比实验：通过全连接神经网络和循环神经网络进行词性标注和命名实体识别，并分别额外加入命名实体和词性先验知识。图 3 至图 6 为全连接神经网络和循环神经网络进行词性标注和命名实体识别的实验结果。

根据结果可以发现，使用全连接神经网络进行词性标注和命名实体识别时，分别加入命名实体和词性先验知识后加快了模型的收敛速度，提升了其准确性。当使用循环神经网络进行命名实体识别时，加入词性先验知识同样加快了模型的收敛速度，提升了其准确性。但当使用循环神经网络进行词性标注时，加入命名实体信息使得模型无法收敛，且其准确率远小于未加入命名实体信息模型的准确率。



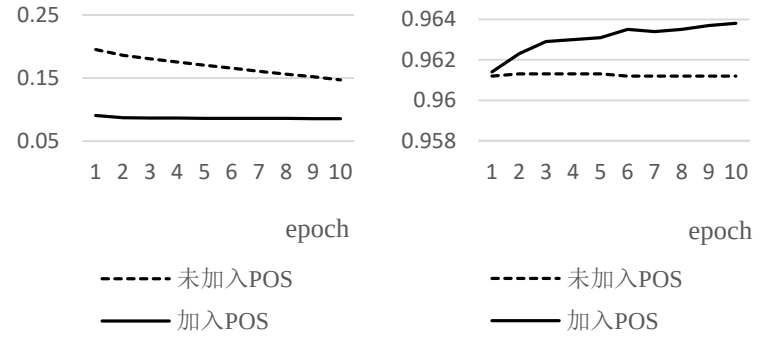
a 损失值 **b 准确率**

图 3 全连接神经网络词性标注



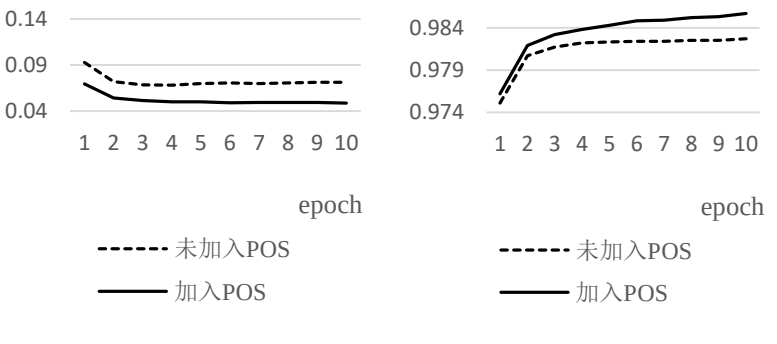
a 损失值 **b 准确率**

图 4 循环神经网络词性标注



a 损失值 **b 准确率**

图 5 全连接神经网络命名实体识别



a 损失值 **b 准确率**

图 6 循环神经网络命名实体识别

3. 2 先验知识对文本分类任务中神经网络模型的影响

3.2.1 实验数据

文本分类实验的数据选自 THUCNews，来自清华大学的中文文本分类工具包 THUCTC[1]。THUCNews 是根据新浪新闻 RSS 订阅频道 2005 至 2011 年间的历史数据筛选过滤生成，其包含了 83 万多篇新闻文档，由人工分成 14 个类别。本研究从中随机抽取 21 万篇新闻作训练集，开发集和测试集各 2000 篇。

3.2.2 数据预处理

我们使用结巴分词对数据进行预处理分词，然后通过斯坦福自然语言处理工具进行词性标注和命名实体识别，构成额外的特征输入。

3.2.3 神经网络模型设计及参数设定

为探究不同的先验知识对于不同的神经网络模型的影响，我们主要采用了两类模型，即基于卷积神经网络模型和基于循环神经网络模型，其模型结构如图 7 和图 8 所示。本研究采用基本的神经网络模型结构，根据语料及开发集设置的具体参数如下。

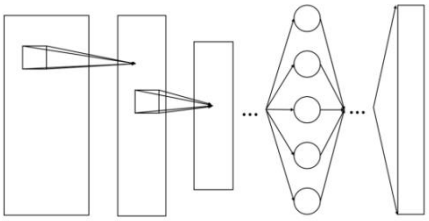


图 7 基于卷积神经网络的文本分类模型

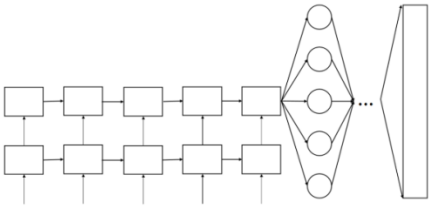


图 8 基于循环神经网络的文本分类模型

基于卷积神经网络的模型结构为：首先以词向量作为输入，其最大长度为 800。通过 2 层卷积神经网络构建文本并挖掘特征，再通过 3 层全连接神经网络做分类，如图 7 所示。其中，词向量的维度为 256，词表大小为 50000。1 层卷积神经网络包含 1 层二维卷积及 1 层二维最大池化层。在第一层和第二层的卷积神经网络中，滤波器的长度分别为 7 和 5，全连接网络的隐藏层大小为 256，且卷积神经网络和全连接神经网络中均加入 dropout 层，其随机断开比为 50%。同时，额外的特征向量紧接词向量，其维度根据词性和命名实体类别数量以及开发集决定，词性向量大小为 40，命名实体向量大小为 20。

基于循环神经网络的模型结构为：首先以词向量作为输入，其最大长度同样为 800。然后通过 2 层循环神经网络挖掘文本特征，再通过 3 层全连接神经网络做分类，如图 8 所示。其中，循环神经网络选择的是长短时记忆神经网络，循环神经网络和全连接网络的隐藏层大

[1]孙茂松，李景阳，郭志芃，赵宇，郑亚斌，司宪策，刘知远. THUCTC：一个高效的中文文本分类工具包. 2016.

小均为 256，词表大小为 50000，并且循环神经网络和全连接神经网络中加入 dropout 层，其随机断开比为 30%。同时，额外的特征向量紧接词向量，词性向量大小为 40，命名实体向量大小为 20。

3.2.4 实验结果与分析

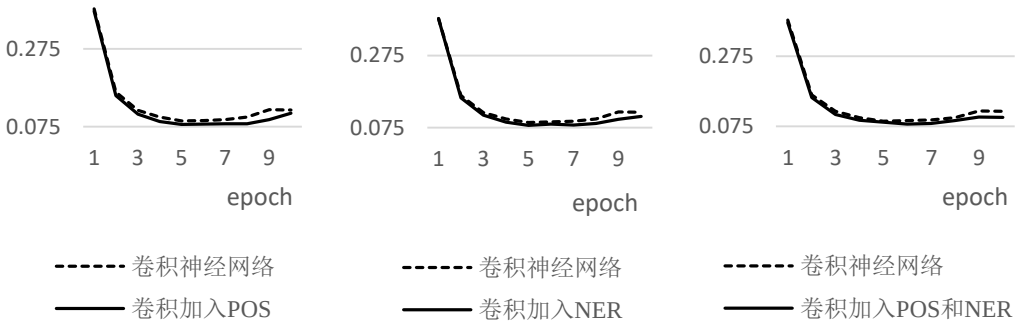
我们分别进行了 2 组实验，使用基于卷积神经网络的模型和基于循环神经网络的模型进行文本分类，其中每组除了基本的神经网络模型，还额外加入了先验知识进行比较，具体实验结果如图 9 至图 12 所示。

根据结果可以发现，在基于卷积神经网络的模型中加入先验知识，加快了模型的收敛速度，提高了模型的准确率。但当同时加入词性和命名实体信息时，结果并未有较大程度的提升，而是和单一加入先验知识的效果接近。而在基于循环神经网络的模型中加入单一先验知识，虽然模型的收敛速度和准确率均有一定程度的提升，但提升效果并不显著。反而在同时加入词性和命名实体后降低了模型的收敛速度和准确率，起到了一定的负面效应。

3. 3 先验知识对机器翻译任务中神经网络模型的影响

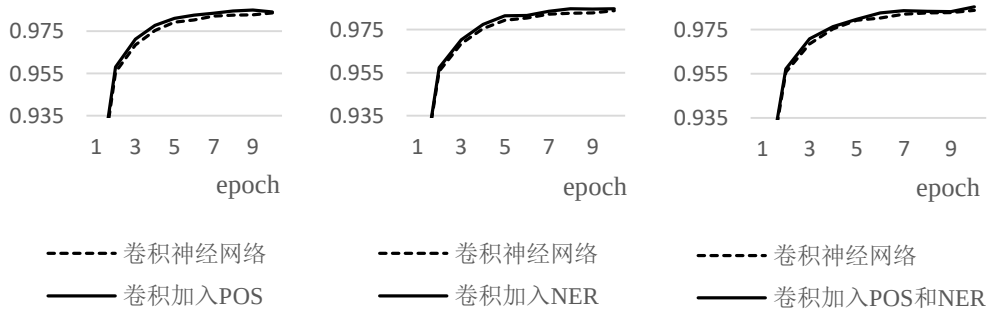
3.3.1 实验数据

神经网络机器翻译实验的数据选自联合国平行语料库 1.0 版，我们采用其中的英中语料，包含 1500 多万平行英中句对，并从中随机挑选出 529 万英中平行句对作为训练集。开发集以及测试集也来自该语料库的英中测试集，各包含 2000 句英中平行句对。



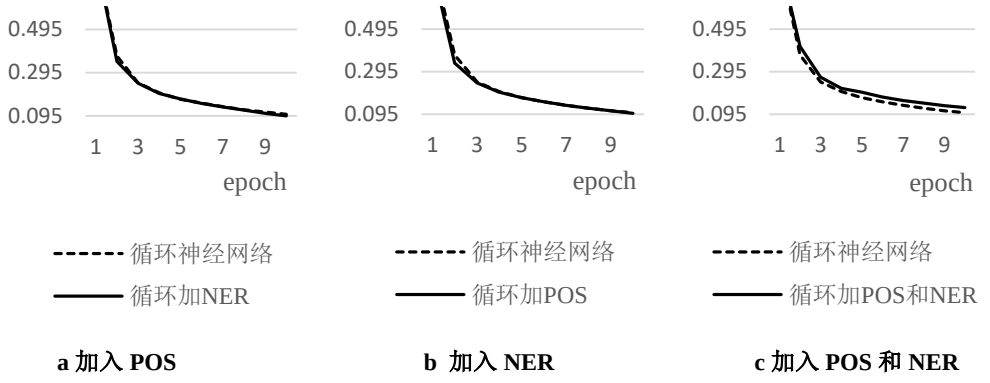
a 加入 POS **b 加入 NER** **c 加入 POS 和 NER**

图 9 卷积神经网络文本分类损失值



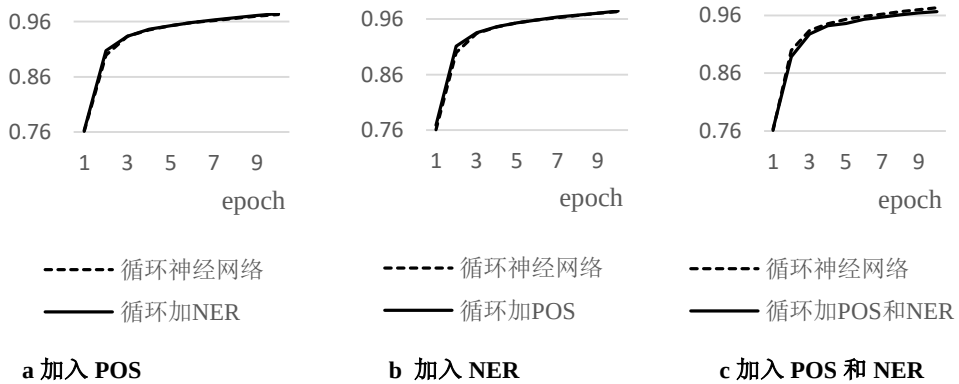
a 加入 POS **b 加入 NER** **c 加入 POS 和 NER**

图 10 卷积神经网络文本分类准确率



a 加入 POS **b 加入 NER** **c 加入 POS 和 NER**

图 11 循环神经网络文本分类损失值



a 加入 POS **b 加入 NER** **c 加入 POS 和 NER**

图 12 循环神经网络文本分类准确率

3.3.2 数据预处理

我们使用 Mosesdecoder 和结巴分词分别对英文和中文进行分词，然后通过斯坦福自然语言处理工具对英文和中文进行词性标注和命名实体识别后构成所需的额外特征输入。考虑到句子长度以及中英词数的比例，我们去除了句子词数超过 50 和词数比例超过 9 的句子，最后得到 460 多万基本符合要求的句对作为最终的训练集。开发集以及测试集均做相同的处理，最后得到 2000 符合要求的英中句对。

3.3.3 神经网络模型设计及参数设定

神经网络机器翻译系统的模型基本还是采用基于注意力机制的编码到解码模型。其中，把源语言的词映射为空间向量作为模型输入层，紧接着使用循环神经网络层作为编码层和解码层，中间通过全连接神经网络实现注意力机制。本研究使用 OpenNMT[2]开源工具来帮助搭建机器翻译系统。

实验中，我们分别使用 2 层循环神经网络作为编码层和解码层，循环神经网络均为长短时记忆网络，可以记忆长期依赖关系。英语和中文的词表大小均为 80000，并且所有的循环神经网络的隐藏层大小均为 800，词向量的维度为 500。使用批随机梯度下降法作为优化方法，其中批的大小为 128，学习速率为 1.0，在第 9 个 epoch 开始衰减为之前学习速率的一半，一共跑了 15 个 epoch。词性标注以及命名实体识别的维度大小分别为 50 和 25，并且接在词向量的后面。以上参数主要根据语料的数量以及开发集来选择。

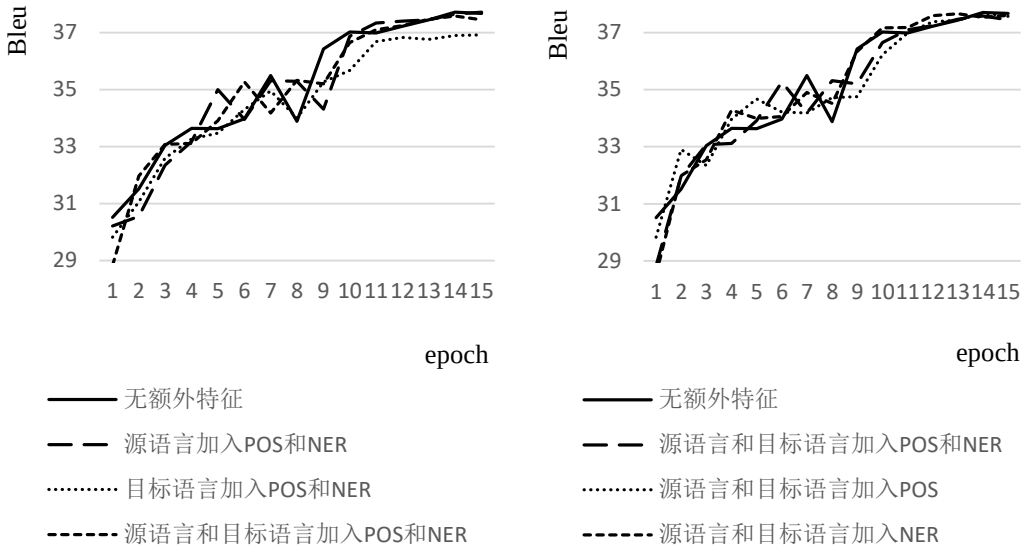
3.3.4 实验结果与分析

我们进行了 2 组对比性实验，分别比较了不同先验知识和不同先验知识的输入位置对神

[2]https://github.com/OpenNMT/OpenNMT

经网络机器翻译的影响。在不同先验知识的实验组中，分别同时在源语言和目标语言中加入词性和命名实体信息；而在不同先验知识输入位置的实验组中，我们分别在源语言和目标语言加入词性和命名实体信息。具体实验结果见图 13 所示。

根据图 13a 可以发现，系统主要通过源语言输入的先验知识获取特征，而对于目标语言位置的先验知识并不能很好的处理，反而增加了其训练的难度，模型收敛也变得更缓慢。同时，可以明显地看出，在源语言输入先验知识和同时在源语言与目标语言输入先验知识的翻译效果与无额外先验知识的系统相比差异并不显著，甚至还需要更多的时间去收敛。根据图 13b 可以发现同时在源语言和目标语言加入不同先验知识与未加入额外的先验知识的翻译效果并没有较明显的差别，甚至收敛速度还会有些许降低，这与图 13a 的观察结果是吻合的。



a 先验知识的输入位置不同 **b 在源语言和目标语言输入不同先验知识**

图 13 先验知识对神经网络机器翻译的影响

4 总结

本研究在自然语言处理基本任务、文本分类和机器翻译三类典型任务中，分别针对不同的语言层先验知识、不同的神经网络结构开展了一系列大规模的实验研究，并且对于机器翻译任务，还比较了不同先验知识对于不同输入位置的影响。根据实验结果，可以发现：不同先验知识对于自然语言处理神经网络模型的影响往往是不同的，先验知识对于不同 NLP 任务和不同神经网络模型的影响也是不同的，不同的先验知识输入位置对于神经网络模型的影响也不同。综合来看，针对不同 NLP 任务和不同神经网络的特点，基于应用场景来挑选不同的先验知识将更加合适。比较不同任务，越复杂的任务，先验知识对模型的影响往往越小，这主要是由于在复杂模型中不一定能有效地获取输入的先验知识，且在本研究的实验中输入的先验知识有限所致。

5 未来工作

接下来，我们还将进一步探寻以下内容：

（1）本研究中加入的先验知识仅包括词性以及命名实体信息，且实验中仅设计了三类任务,更多的先验知识以及更丰富的自然语言处理任务尚有待验证，本研究得出的观察结论的泛化性也还有待进一步拓展验证。

（2）本研究中标注的词性和命名实体信息均来自于斯坦福大学的自然语言处理标注工

具，可能存在一定的工具使用误差，如何减少前期误差对神经网络模型的影响有待探索。

（3）本研究中加入词性和命名实体信息的方法较简单和直接，仅仅使用了与词向量拼接的方式。如何进一步生成有先验知识的词向量后以不同的方式加入先验知识仍有待探索。

（4）在机器翻译任务中，由于时间有限，当前模型只训练到第 15 个 epoch。加入先验知识的模型由于加入了更多的参数，往往需要更多的训练来提取输入的有效先验知识。如果继续训练，是否可以在翻译效果上显著超过未加入先验知识的模型还需进一步尝试。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[2] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003: 173-180.

[3] Shen H, Sarkar A. Voting between multiple data representations for text chunking[C]// Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence. Springer Berlin Heidelberg, 2005: 389-400.

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

[5] Koomen P, Punyakanok V, Roth D, et al. Generalized inference with multiple semantic role labeling systems[C]//Proceedings of the Ninth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2005: 181-184.

[6] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(Aug): 2493-2537.

[7] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C]//EMNLP 2013: 647-657.

[8] Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[R]. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[9] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[10] Lai S, Xu L, Liu K, et al. Recurrent Convolutional Neural Networks for Text Classification[C]//AAAI 2015, 2267-2273.

[11] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. 2015: 649-657.

[12] Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003: 48-54.

[13] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[14] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

作者简介：

贝超（1992-），男，硕士，主要研究领域为自然语言处理。Email: beichao202@163.com.

胡珀（1980-），男，博士，副教授，通讯作者，主要研究方向为自然语言处理、自动文摘。Email: phu@mail.ccnu.edu.cn.