

# 实体驱动的双向 LSTM 篇章连贯性建模\*

杜舒静

徐凡

王明文<sup>+</sup>

shujingdu@jxnu.edu.cn

fanxu@jxnu.edu.cn

mwwang@jxnu.edu.cn

江西师范大学计算机信息工程学院, 江西, 南昌, 33002

**摘要:** 篇章连贯性建模是自然语言处理研究领域的一个基础问题。主流的篇章连贯性模型分为两大类, 分别是基于实体网络的连贯性模型和基于神经网络的篇章连贯性模型。其中, 基于实体网络的篇章连贯性模型需要进行特征提取, 而基于深度学习的模型没有充分考虑到篇章中句子间的实体链接对连贯性建模的重要作用。基于此, 本文首先抽取篇章中相邻句子的实体信息, 将其进行分布式表示, 然后将此信息通过多种简单且有效的向量操作融合至句子级的双向 LSTM 深度学习模型之中。在汉语和英语篇章语料上的句子排序和中英文机器翻译连贯性检测两种任务上的实验表明本文提出的模型性能和现有模型相比有所提升, 尤其在中文上有显著提升。

**关键词:** 实体; 篇章连贯性; 双向 LSTM; 分布式表示

**中图分类号:** TP391

**文献标识码:** A

## An Entity-driven Bi-directional LSTM-based Model for Discourse Coherence Modeling

Shujing Du

FanXu

Mingwen Wang<sup>+</sup>

shujingdu@jxnu.edu.cn

fanxu@jxnu.edu.cn

mwwang@jxnu.edu.cn

School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China

+ Corresponding author

**Abstract:** Discourse coherence modeling is a fundamental problem in natural language processing. Mainstream coherence model have two kinds, The entity based coherence model and deep learning based model. The entity based coherence model need to exact features, while the deep learning based model do not take full consideration of using the important information, which is the entity link among sentences across discourse. Based on this, this paper firstly extracts the entity information on the adjacent sentences within the discourse, then represent it in a distributed way, and finally integrate them into the sentence-level bidirectional LSTM model through various simple vector operation. The experiment on Chinese and English sentence ordering task and Chinese-English statistical machine translation coherence detection task shows that our model outperforms the state-of-the-art ones, especially in Chinese.

**Keyword:** Entity; Discourse Coherence; Bi-directional LSTM; Distributed representation

---

\*基金资助: 国家自然科学基金(61402208,61462045), 江西省自然科学基金(20151BAB207027), 江西省研究生创新基金(YC2016S169), 国家语委"十二五"科研规划项目(YB125-99)。

## 1 引言

篇章连贯性是衡量篇章中句子间语义是否通顺和形式是否衔接的一个重要因素。因此，篇章连贯性建模对于统计机器翻译、自然语言问答、自然语言生成具有重要的作用。根据系统功能语言学家 Halliday 所述，连贯性是一种具有逻辑的内在语义关联，连贯的篇章是由一些逻辑或语义上存在某种相似元素的段落或句子组成，是篇章的一个基本特征<sup>[1]</sup>。此外，衔接性是篇章的另外一个基本特性，它是连句成章的词汇和语法方面的手段，它是指语篇中表层结构上的黏着性，是语篇的有形网络；相比较而言，连贯是采用这些手段所产生的结果，它指语篇中底层语义上的关联性，是语篇的无形网络<sup>[2]</sup>。下面例子分别为连贯和不连贯的篇章。

### 例 1：连贯的篇章实例。

① 张三吃了很多食物，②由于所吃的食物大部分是高脂肪的，③所以张三体重超标了。

### 例 2：不连贯的篇章实例。

① 张三吃了很多食物，②由于他昨天买了很多衣服，③所以张三体重超标了。

其中，例 1 的分句①和分句②阐述了事件的主人公及发生的事件，分句③说明了事件的结果，三个分句都存在相似的成分“张三”；然而例 2 中的分句②和其余的分句在内容上是不相关的，其中分句①和分句③是在阐述事件的主人公以及事件的结果，而分句②却在阐述一个与当前篇章无关的事件。从篇章的衔接性（实体重复）角度看，例 1 是连贯的篇章，而例 2 则是不连贯的篇章，因此，衔接性（实体重复）有助于篇章连贯性建模。

篇章连贯性建模是自然语言处理中的一个基础问题，它在多语言即时机器翻译<sup>[3]</sup>、文本自动摘要<sup>[4][5][6]</sup>、作文自动评分<sup>[7][8][9]</sup>等方面均有广泛的应用前景。传统的基于实体网络的篇章连贯性模型需要花费大量的时间进行特征提取；且太依赖实体特征，如果篇章中含有的实体较多，可能会对实体网络模型产生一些噪音，导致实体网络模型的准确率下降。而现有的基于深度学习的模型仅学习了文本本身的信息来对文本进行篇章连贯性建模，没有考虑到篇章中句子间的实体链接对于篇章连贯性建模的重要性作用，从而只是初步学习了篇章连贯性来对其进行建模，并未能进一步挖掘和利用篇章的连贯性。基于此，本文模拟人阅读篇章时，利用篇章连贯性的名词重复阅读习惯进一步探究篇章连贯性。本文首先抽取出篇章中相邻句子的实体信息，将其进行分布式表示来模拟人利用名词重复阅读的习惯，然后将此信息通过向量操作融合至句子级的双向 LSTM 深度学习模型之中。在汉语和英语篇章语料上的句子排序和机器翻译连贯性检测两种任务上的实验表明本文提出的模型性能较现有模型具有显著提升。

本文的后续内容如下：第 2 节将介绍了篇章连贯性建模的相关工作；第 3 节着重阐述了本文提出的实体驱动的双向 LSTM 篇章连贯性模型；第 4 节给出了实验设置及详细的实验结果分析；第 5 节是结论及后续工作部分。

## 2 相关工作

本节主要从基于特征抽取的篇章连贯性模型和基于深度学习的连贯性篇章模型两个方面进行概述相关工作。

### 2.1 基于特征抽取的模型

传统的篇章连贯性建模主要通过抽取反映篇章连贯性方面的特征（如：文本的潜在语义、实体的语法成份或篇章修辞关系等）进行工作。其中，1998年 Foltz P W.<sup>[10]</sup>提出了基于潜在语义(Latent Semantic Analysis, 简称 LSA)的篇章连贯性模型, 该模型通过学习文本的潜在语义对文本的篇章连贯性建模, 获取文本的潜在语义。但是, 该模型在降维的时候可能会丢失一些包含篇章连贯的信息, 导致无法捕获较为全面的篇章连贯信息。2004年 Barzilay.<sup>[4]</sup>提出了基于隐马尔科夫(Hidden Markov Model, 简称 HMM)的篇章连贯性模型, 该模型通过对文本的主题以及这些主题出现的顺序进行分析, 对特定领域内的文本内容结构进行篇章连贯性建模。该模型的局限是受到领域的影响很大, 不适用于开放的领域。此外, 基于实体的网格模型<sup>[4][11][12][13]</sup>是目前较为流行的传统模型, 该模型通过将篇章中实体作为网格的格点, 将不同句子间实体的语义角色(如: 主语、谓语和其它等)的转换作为格点间的连线, 以网格的形式描绘对篇章的连贯性加以建模。此后的很多工作都是基于该模型开展的改进工作<sup>[14][15][16]</sup>。综上所述, 传统的篇章连贯性建模均需要进行提取特征工作, 代价总体而言比较高。

### 2.2 基于神经网络的模型

当前, 深度学习在图像处理<sup>[17]</sup>、计算机视觉<sup>[18]</sup>和语音识别<sup>[19]</sup>等方面均取得了非常显著的成绩。同样, 其在自然语言处理的统计机器翻译<sup>[20]</sup>、问答系统<sup>[21]</sup>、自然语言生成<sup>[22]</sup>等方面也取得了不错的成果。针对篇章连贯性建模方面, 2014年李纪为等人提出基于分布式句子向量表示的神经网络模型<sup>[23]</sup>, 该模型分别利用两种 RNN (Recurrent Neural NetWork 和 Recursive Neural NetWork ) 表示篇章中的句子分布式向量表示, 然后利用构建团的方式对篇章连贯性加以建模。此外, 2016年李纪为等人提出了开放领域的生成式模型和判别式模型<sup>[24]</sup>, 该模型通过学习不连贯的文本以及训练大量的开放领域语料来获取篇章连贯性。综述, 基于神经网络的模型可以免去传统方法中繁琐的特征工程步骤, 自动地提取出篇章中表示连贯性方面的各种词汇和句法特征, 具有一定的优势。

## 3 实体驱动的深度学习方法

递归神经网络(Recursive Neural Networks, RNNs)能够有效地对时间关系进行建模, 通过输入序列到隐藏层的映射能够学习到复杂的时间动态关系, 已被成功应用于语音识别、机器翻译和行为识别等领域, 尤其是时间递归神经网络(Long-Short Term Memory, 简称 LSTM), 能够应用于处理和预测时间序列中间隔和延迟非常长的重要事件。基于此, 本文

将利用双向 LSTM 构建篇章中句子的分布式向量表示形式。

由于现有的基于深度学习的篇章连贯性模型没有充分考虑到篇章中句子间的实体链接对于篇章连贯性建模的重要性作用。基于此，本文首先抽取出篇章中相邻句子间的实体（名词）信息，将其进行分布式表示来模拟句子间实体的链接，然后将此信息通过向量操作融合至句子级的双向 LSTM 深度学习模型之中。为清晰起见，图 1 是本文提出模型的框架。它是由句子的双向 LSTM 表示、句子的实体驱动向量、基于团的篇章连贯性表示三个模块构成。分别如下所示：

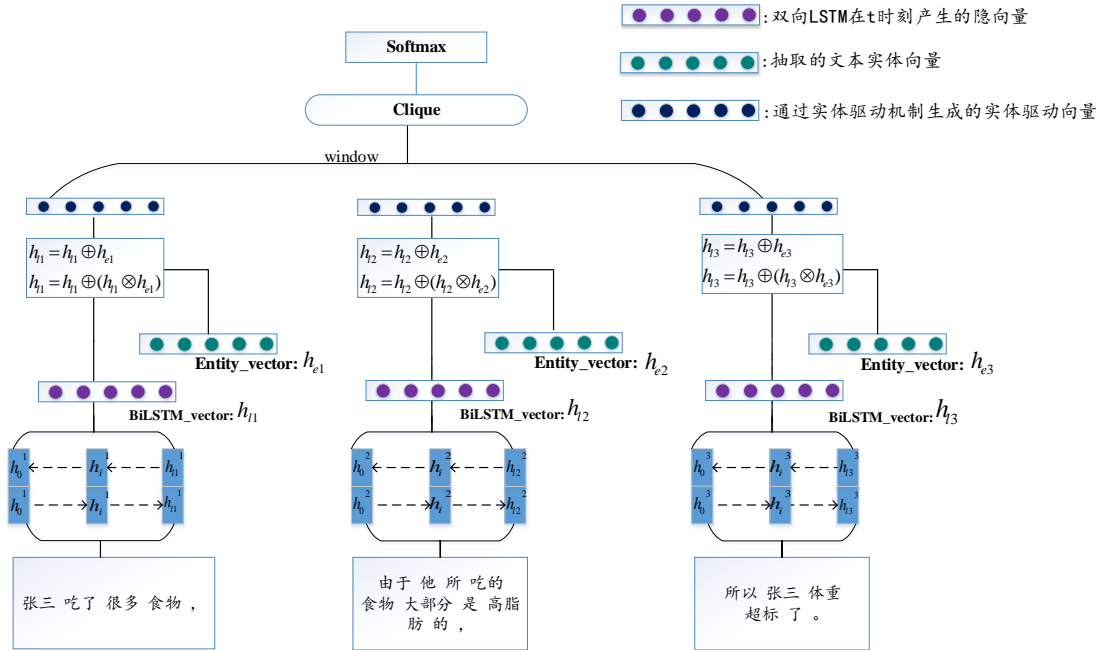


图 1. 实体驱动的双向 LSTM 模型

从图中可以看出本文的模型首先借助分词工具对语料进行分词，之后用词向量工具训练语料生成词向量，将得到词向量输入 lstm 中得到句子的表示 BLSTM\_vector，同时本文将分词后的语料进行词性标注，并且语料中名词对应的向量 Entity\_vector 提取出来，并将 BLSTM\_vector 和 Entity\_vector 进行图中的公式的运算操作，得到本文提出的实体驱动向量。之后将实体驱动向量融入相邻一定大小句子组成的团，就句子排序任务而言，每次随机替换原文中的句子生成负例，原始文本为正则；如果置换文本的连贯性得分大于原始文本的得分，则将其看成连贯的文本，否则是不连贯的文本；就机器翻译连贯性任务评估而言，每次学习人工译文为正则，机器生成译文为负例，计算和判别过程类似。本文将在接下来的 3.1-3.3 小节自底向上的描述图 1 的组成。

### 3.1 句子的双向 LSTM 表示

首先，本文采用 glove<sup>1</sup>和 word2vector<sup>2</sup>训练篇章语料生成词向量表示形式，可以为句子中的每个词  $w$  训练生成一个  $k$  维向量表示  $x_w \in \mathbb{R}^k$ 。语料中的每个句子都可以看成一系列词向量，用  $x_i$  表示句子  $S$  中的第  $i$  个词向量，则含有  $L$  个词的句子可以被表示为  $S_L = [x_1, x_2, \dots, x_L]$ ，最终通过双向 LSTM 训练得到  $k$  维句子向量  $S_L$ 。图 1 中 LSTM 模型的相关

<sup>1</sup> <http://code.google.com/p/word2vec/>

<sup>2</sup> <http://nlp.stanford.edu/projects/glove/>

参数计算公式如下：

$$i_i = \text{sigmoid}(W_i[x_i, h_{i-1}] + b_i) \quad (1)$$

$$\theta_i = \tanh(W_c[x_i, h_{i-1}] + b_c) \quad (2)$$

$$f_i = \text{sigmoid}(W_f[x_i, h_{i-1}] + b_f) \quad (3)$$

$$o_i = \text{sigmoid}(W_o[x_i, h_{i-1}] + b_o) \quad (4)$$

$$c_i = i_i * \theta_i + f_i * c_{i-1} \quad (5)$$

$$h_i = o_i * \tanh(c_i) \quad (6)$$

根据公式(1-6)的定义可知)，与单一的  $\tanh$  循环体不同，LSTM<sup>[25][26][27]</sup>有三个影响细胞状态的“门”结构，分别是：输入门、遗忘门和输出门。LSTM 靠这些“门”结构让信息有选择地影响循环神经网络中的细胞状态，所谓的门结构就是利用 sigmoid 神经网络和一个按位做乘法的操作来实现的。其中输入门和遗忘门是 LSTM 的核心，遗忘门会根据当前的输入  $x_i$ 、上一时刻细胞状态  $c_{i-1}$  和上一时刻输出  $h_i$  共同决定哪一部分记忆需要忘记；输出门会根据最新的细胞状态  $c_i$ ，上一时刻的输出  $h_{i-1}$  和当前时刻的输入  $x_i$  来决定该时刻的输出  $h_i$ 。文采用双向的 LSTM 得到的篇章中的句子的向量表示。如图 1 所示用一个双向 LSTM 神经网络，本文可以得到第  $n$  个词的上下文信息  $h_n = [\overset{r}{h}_n, \overset{s}{h}_n]$ ， $\overset{r}{h}_n, \overset{s}{h}_n \in \mathbb{R}^d$ ，之后通过拼接词的上下文信息来表示句子。

### 3.2 句子的实体驱动向量

借助双向 LSTM 获得了文本中句子的分布式表示后，本文希望模拟人阅读文章时利用名词重复阅读的习惯进行篇章连贯性建模。即通读文章时，人们往往会记住文章中相邻句子中的一些实体（名词），然后根据这些实体间的重复关系来推测整篇文章的大义。鉴于此，本文设计了实体驱动的方法来强化篇章中句子间的名词的重要作用。

本文首先将文本中句子间的实体（名词）抽取出来，然后得到它们的分布式向量表示形式，最后将这些实体向量(见公式 7 所示，其中符号  $\oplus$  代表向量点加操作)与句子的双向 LSTM 分布式表示进行拼接,得到句子级别的最终向量表示(见公式 8 和 9 所示，其中符号  $\oplus$  代表向量点加操作，符号  $\otimes$  代表向量点乘操作)：

$$H_{entity} = w_{NN1} \oplus w_{NN2} \oplus \dots \oplus w_{NNL} \quad (7)$$

对于原始句子的表示，本模型首先将语料借助词向量工具训练得到的词向量输入双向 LSTM，得到隐向量  $h_i$ (其初始值是将训练好的词向量输入双向 LSTM 后得到的句子向量)，同时用一个额外的存储容器存储文本中的实体向量。之后受到马尔可夫链的启发，即第  $n$  个词的状态只与第  $n-1$  个词的状态有关。最初本文对实体向量仅仅进行了点加操作和点乘操作，来得到实体向量的多种表示方式，之后本文对数据集和实验结果进行分析，发现有一些

句子可能不存在实体向量，仅对向量进行点乘的操作可能太粗糙了，因此本文对点乘操作进行了类平滑操作得到了本文的乘法操作。本文采用的加法和乘法操作（见公式 8 和公式 9）：

:

$$H_t < -H_t \oplus H_{e_{n_t}} \quad (8)$$

$$H_t < -H_t \oplus (H_t \otimes H_{entity}) \quad (9)$$

### 3.3 基于团的篇章连贯性表示

本模型首先将含有  $L$  个相邻句子的滑动窗口视为团  $C$ ，利用团  $C$  中所有的句子向量拼接表示团向量，其大小为  $(L+1) * k$  维：

$$L(C) = \tanh(W \underline{[x_{sn}, \dots, x_{sn+L/2}, \dots, x_{sn+L}]}) \quad (10)$$

然后为每个团  $C$  定义一个相应的标签  $y_c$ ，如果团  $C$  是连贯的，则标签  $y_c$  的值为 1，否则标签  $y_c$  的值是 0。同时，将团  $C$  向量输入 sigmoid 函数（见公式 11）：

$$p(y_c = 1) = \text{sigmoid } U^T q_c + b \quad (11)$$

最后，整个篇章的连贯性得分为每个团的连贯性得分乘积（见公式 12），本文认为连贯性得分越大的文本越连贯：

$$S = \prod p(y_c = 1) \quad (12)$$

即假设两篇文本  $d_1$  和  $d_2$  的连贯性得分分别为  $S_{d_1}$  和  $S_{d_2}$ ，如果  $S_{d_1}$  大于  $S_{d_2}$ ，则文本  $d_1$  比文本  $d_2$  更连贯。

### 3.4 模型训练和优化

对于句子排序任务。本文首先用训练集来训练本文提出的模型，即将人工生成的文本作为正例，随机置换原始文本的句子生成的文档作为负例，本文的模型要学到测试集中的正例和负例的连贯性得分之间存在的关系；之后再用本文的模型对测试集进行打分得到句子排序的连贯性分数。对于机器翻译连贯性评估。本文用人工译文作为训练集的正例，机器翻译得到的译文作为负例。首先用本文的模型学到测试集中的正例和负例的连贯性得分之间存在的关系；之后再用本文的模型对测试集进行打分得到机器翻译连贯性评估分数。

本模型采用目前广泛使用的交叉熵函数作为目标函数（见公式 13）。

$$J(\Theta) = \frac{1}{M} \sum_{C \in \text{trainset}} \{-y_c \log[p(y_c = 1)] - (1 - y_c) \log[1 - p(y_c = 1)]\} + \frac{Q}{2M} \sum_{\theta \in \Theta} \theta^2 \quad (13)$$

其中， $M$  代表训练实例的个数； $\Theta$  是模型参数； $Q$  作为正则化参数，用来防止模型过拟合。 $g_t^i$  表示参数  $\theta_t^i$  的第  $t$  步反向传播的梯度，则在第  $t$  步的参数更新（见公式 14）：

$$\theta_T = \theta_{T-1} - \frac{\alpha}{\sum_{t=0}^T \sqrt{g_t^2}} g_T^i \quad (14)$$

本文采用了自适应梯度下降法 AdaGrad<sup>[28]</sup>来优化此损失函数，它能够对每个参数自适应不同的学习速率，对稀疏特征，得到大的学习更新，对非稀疏特征，得到较小的学习更新，是目前广泛使用的优化算法。

## 4 实验

为了验证本文提出的实体驱动双向 LSTM 篇章连贯性模型的性能，本文在标准语料库上进行了汉语和英语环境下的句子排序和机器翻译连贯性评估两组实验。其中，句子排序任务实验是给定原文本和随机置换句子顺序的置换文本，假定原文本比置换文本更连贯；机器翻译连贯性评估则是自动识别出比机器自动生成译文更连贯的人工参考译文。同时，本文进行了 5 倍交叉实验对模型加以验证。

### 4.1 数据集及评价指标

**英文句子排序数据集：**本文采用了文献<sup>[12]</sup>中使用的两种不同文体的国际基准语料：地震语料(earthquake)和飞机失事(airplane)语料。其中，训练集分别由 99 篇地震语料和 100 篇飞机失事语料组成，测试集分别由 100 篇地震语料和 100 篇飞机失事语料以及随机置换测试原始文本中句子顺序生成的置换语料组成，同时规定每篇原始文本最多生成 20 个置换文本。

**中文句子排序数据集：**本文分别采用了两种汉语篇章数据集作为语料。其一是汉语树库 CTB(Chinese TreeBank)6.0 对应的原始文件，其二是本文学校某学院标注的 130 篇汉语作文。其中，对于 CTB 语料，本文选择 chtb\_2946 到 chtb\_3045 在内的 100 个文档作为训练数据，chtsb\_3046 到 chtb\_3145 在内的 99 个文档及其置换文本作为测试数据集。对于汉语作文语料（汉语作文），本文选取其中的前 65 篇作为训练集，其余的文档以及随机置换文档作为测试集，同时规定每篇原始文本最多生成 20 个置换文本。

**机器翻译连贯性评估数据集：**本文从语言数据联盟（Linguistic Data Consortium，简称 LDC）中提取 NIST Open Machine Translation 2008 评估中的汉英翻译作为语料 (Smt\_translation)，其的索引号为 LDC2010T01 和 ISBN1-58563-533-2 其中该数据集中含有 128 组成对的文件（128 篇人工参考译文和 128 篇机器自动翻译得到的译文）。

**语料预处理：**本实验的词向量是利用词向量工具 word2vector<sup>[29]</sup>以及 glove<sup>[30]</sup>训练得到；同时采用斯坦福的词性标记工具<sup>3</sup>提取出篇章中句子的实体，同时利用 ICTCLAS<sup>4</sup>进行中文分词。

**评价指标：**本实验采用的评价指标是准确率，即数据集中有  $m$  篇文本，预测数据集中连贯的文本数为  $n$ ，准确率等于  $n$  除以  $m$  得到的数值。

此外，本文参考文献[18]的参数设置，将 AdaGrad 的初始学习率设置为 0.1，词向量的维数设置为 300，防止过拟合的 drop rate 设置为 0.2，滑动窗口的大小设置为 5。

### 4.2 实验结果

本节分别给出汉语和英语环境下的句子排序任务，以及机器翻译连贯性评估任务下的实验结果，并进行详细的实验结果分析。

---

<sup>3</sup> <http://nlp.stanford.edu/software/>

<sup>4</sup> <http://ictclas.nlpir.org/downloads>

#### 4.2.1 中英文句子排序实验结果

为了体现英汉实验环境的公平性，本文首先采用英文语料重现了英文语料下的神经网络实验，然后将其移植至中文环境。同时，本文分别验证了 word2vec 和 glove 两种词向量下的性能，由于 word2vec 显著优于 glove 性能，为了找到最适合做篇章连贯性建模的语料设置，本文首先采用 word2vec 和 glove 生成的不同维数的向量进行实验；同时为了找出较适合篇章连贯性的语料形态，在语料预处理时本文将语料生成为一句一行及一词一行的形式，以期望找到较适合篇章连贯性建模的语料形态。通过对该实验的结果进行分析，本文发现形态为一句一行的语料，借助 word2vec 生成的 300 维词向量最适合篇章连贯性建模，限于篇幅，本文以下均提供采用该种词嵌入的实验结果。

本文采用的基准模型如下：

**基准模型 1：**实体网络模型(Entity graph based model<sup>[14]</sup>)，是一个代表性的基于特征工程的传统篇章连贯性模型，本文利用该文献中提供的开源代码<sup>5</sup>重现了该模型，同时将此模型移植至中文环境。

**基准模型 2：**神经网络模型(Li et al<sup>[24]</sup>)，该模型仅利用双向 LSTM 对篇章连贯性进行建模，本文重现了此模型，同时将其移植至汉语环境。

表 2. 中英文句子排序性能

Model	英文		中文	
	Earthquake	Airplane	ctb	汉语作文
Entity graph based model <sup>[14]</sup>	0.7119	0.7863	0.6453	0.6886
Li et al <sup>[24]</sup>	0.9951	0.8775	0.6461	0.7703
本文的模型（相加操作）	<b>0.9976</b>	0.8585	0.6895	<b>0.8363*</b>
本文的模型（相乘操作）	0.9962	<b>0.8866</b>	<b>0.7063*</b>	0.8350*

表 2 显示了中英文句子排序任务的系统性能对比实验结果。该实验采用的语料是 4.1 节中的中文句子排序数据集和英文句子排序数据集，该实验结果表明本文提出的实体驱动模型明显优于基准模型。原因在于 airplane 语料中有约 29.32% 的名词，earthquake 语料中有约 30.67% 的名词，作文语料中有约 22.01% 的名词，ctb 语料中有约 25.76% 的名词。对于名词的含量比较大的数据集（类似于 earthquake 和 ctb）而言，由于含有的名词数量较多，对实体网络模型可能存在一些噪声，因而它们在实体网络模型上的结果较低，同时由于一般的深度学习模型没有充分挖掘出这种名词间的依赖关系，导致模型的性能也存在一定的局限。根据进一步分析，本文的模型(相加操作)更适合学习名词较多的语料集，本文的模型（相乘操作）则更适合学习名词较少的语料集。其中，本文的模型(相加操作)主要是对抽取的实体向量与 LSTM 向量进行按位相加操作，而本文的模型(相乘操作)主要是对抽取的实体向量与

<sup>5</sup> <http://github.com/karins/CoherenceFramework>.



LSTM 向量进行按位相乘操作。总体上本文的模型(相乘操作)比较稳定，显著优于现有的基于实体和基于神经网络的模型。结果表明本文提出的实体向量对于篇章连贯性建模具有重要作用，而且可以有多种有效的方式（相加或相乘）与现有神经网络模型加以融合。此外，经过显著性测试，本文模型在中文情况较比现有模型具有显著性提升（采用成对的 t-检测对应的 $P<0.01$ ），但对于性能比较好的英文情况，本文模型的性能仍有小幅度的提升。

#### 4.2.2 机器翻译连贯性评估实验结果

表 3 显示了机器翻译连贯性评估任务的性能对比实验结果。该实验采用的数据集是 4.1 节中的机器翻译连贯性评估数据集，该实验结果表明本文的模型（相加操作）更适用于机器翻译的连贯性评估，机器翻译语料中名词有约 24.21%，参考译文中含有约 25.51% 的名词，机器翻译得到的译文中含有约 22.75% 的名词，虽然 Smt\_translation 语料中名词比例与句子排序的中文语料 ctb 和作文语料比例类似，但由于机器翻译连贯性检测任务比句子排序任务更具有挑战性，其原因在于前者需要区分比较相近的人工参考译文与机器生成译文，而句子排序任务中原文件和置换文件的可比较性不高，从而导致传统的实体网格方法和传统神经网络方法的识别性能均不高，但本文将两者融合后，性能得以显著提升。说明本文的模型将实体信息融入机器学习中能较好的评估机器翻译连贯性，充分体现了实体的分布式表示与现有深度学习模型之间具有一定的互补作用。此外，经过显著性测试，显示本文模型在中文情况较现有模型的性能具有显著性提升（采用成对的 t-检测对应的 $P<0.01$ ）。

表 3. 机器翻译连贯性评估五倍交叉验证实验

Model	Smt_translation
Entity graph based model <sup>[14]</sup>	0.4560
Li et al <sup>[24]</sup>	0.4900
本文的模型（相加操作）	<b>0.6492*</b>
本文的模型（相乘操作）	0.6367*

## 5 总结与展望

本文提出了一个实体驱动的神经网络篇章连贯性模型。由实验结果可知本文的模型（相乘操作）更好，在句子排序和机器翻译连贯性评估任务上表现均有稳定的提升。本文的模型既避免了需要耗费大量时间的特征工程，又充分考虑了篇章中句子间的实体链接对于篇章连贯性建模的重要性作用。因此，本文首先抽取出篇章中相邻句子的实体信息，将其进行分布式表示，然后将此信息通过多种简单且有效的向量操作融合至句子级的双向 LSTM 深度学习模型之中。在汉语和英语篇章语料上的句子排序和机器翻译连贯性检测两种任务上的实验结果表明本文提出的模型性能较比现有模型具有显著提升。

## 致谢

在此，本文向对本研究工作提供帮助的老师 and 同学表示感谢，同时对本文的评审专家及提出的宝贵意见表示感谢。

## 参考文献

- [1] Halliday M A K. An Introduction to Functional Gram—mar[M]. New York: Oxford University Press Inc, 2004.
- [2] 黄国文. 语篇分析概要[M]. 长沙: 湖南教育出版社, 1987.
- [3] Heidi J. Fox. Phrasal Cohesion and Statistical Machine Translation[C]// In Proceedings of EMNLP, 2002, pages 304-311.
- [4] Regina Barzilay and Lillian Lee. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization[C]// In Proceedings of NAACL-HLT, 2004, pages 113-120.
- [5] Zi-Heng Lin, Hwee Tou Ng and Min-Yen Kan. Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation[C]// In Proceedings of ACL, 2012, pages 1006–1014.
- [6] Danushka Bollegala, Naoaki Okazaki and Mitsuru Ishizuka. A Bottom-Up Approach to Sentence Ordering for Multi-Document Summarization[C]//In Proceedings of ICCL-ACL, 2012, pages 385-392.
- [7] Helen Yannakoudakis and Ted Briscoe. Modeling coherence in ESOL learner texts[C]// In Proceedings of ACL, 2013, pages 33-43.
- [8] Jill Burstein, Joel Tetreault and Slava Andreyev. Using Entity-Based Features to Model Coherence in Student Essays[C]//In Proceedings of NAACL-HLT, 2010, pages 681-684.
- [9] Derrick Higgins, Jill Burstin, Daniel Marcu, et al. Evaluating Multiple Aspects of Coherence in Student Essays[C]//In Proceedings of NAACL-HLT, 2004, pages 185-192.
- [10] Foltz P W, Walter K, Thomas K L. The measure—ment of textual coherence with latent semantic analy—sis[J]. Discourse Processes, 1998, 25(2&3): 285—307.
- [11] Regina Barzilay and Mirella Lapata. Modeling Local Coherence: An Entity-Based Approach[C]// In Proceedings of ACL, 2005, pages 141-148.
- [12] Regina Barzilay and Mirella Lapata. Modeling Local Coherence: An Entity-Based Approach [J]. Computational Linguistics, 2008, 34(1):1-34.
- [13] Mirella Lapata and Regina Barzilay. Automatic Evaluation of Text Coherence: Models and Representations[C]//In Proceedings of IJCAI, 2015, pages 1085-1090.
- [14] Camille Guinaudeau and Michael Strube. Graph-based local coherence modeling[C]//In Proceedings of ACL, 2013, pages 93–103.
- [15] Feng V W, Hirst G. Extending the entity—based coherence model with multiple ranks[C]//In Proceedings of EACL, 2012, pages 315—324.
- [16] AK John , LD Caro , G Boella. Text Segmentation with Topic Modeling and Entity Coherence[C] //International Conference on Hybrid Intelligent Systems, 2016:175-185.
- [17] Z Wang , J Lu , R Lin , et al. Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images[J]. IEEE Journal of Biomedical & Health Informatics, 2017, PP (99) :1-1.
- [18] A Paul , A Dey, DP Mukherjee, et al. Regenerative Random Forest with Automatic Feature Selection to Detect Mitosis in Histopathological Breast Cancer Images [C]//Springer

International Publishing , 2015: 94-102.

- [19] 杨明翰 , 许曜麒 , 洪孝宗, 等等. 融合多任务学习类神经网络声学模型训练于会议语音辨识之研究[C]// Proceedings of the 28th Conference on Computational Linguistics and Speech Processing, 2016:016-1001
- [20] Haitao Mi, Zhiguo Wang, Abe Ittycheriah. Supervised Attentions for Neural Machine Translation[C]//Proceeding of the EMNLP,2016:2283-2288
- [21] Ming Tan, Cicero dos Santos, Bing Xiang , et al. Improved Representation Learning for Question Answer Matching[C] //In Proceedings of ACL, 2016, pages 464-473.
- [22] Natsuda Laokulrat, Sang Phan, Noriki Nishida, et al . Generating Video Description using Sequence-to-sequence Model with Temporal Attention[C]//In Proceedings of COLING , 2016, pages 44-52.
- [23] Jiwei Li and Eduard Hovy. “A Model of Coherence Based on Distributed Sentence Representation”[C]// In Proceedings of EMNLP, 2014. pages 2039-2048.
- [24] J. Li and D. Jurafsky. Neural net models for open-domain discourse coherence. arXiv preprint arXiv:1606.01545, 2016.
- [25] A Graves. Long Short Term Memory[J]. Springer Berlin Heidelberg, 2012, 9 (8) :1735-1780.
- [26] Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[J]. Computer Science, 2014:338-342.
- [27] Tai, K. S., Socher, R., & Manning, C. D. Improved semantic representations from tree-structured long short-term memory networks[J].Computer Science, 2015, 5(1):26- 36.
- [28] John Duchi, Elad Hazan and Yoram Singer, Adaptive subgradient methods for online learning and stochastic optimization[J], Journal of Machine Learning Research, 2011, 12 (7) :2121-2159.
- [29] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[C] //Proceedings of ICLR Workshop, 2013:3-1127.
- [30] Jeffrey Pennington, Richard Socher, and Christopher D.Manning. 2014. Glove: Global vectors for word representation[C]//In Proceedings of EMNLP, 2014:1523-1543.