文章编号：1003-0077(2020)05-0001-09

基于局部语义相关性的定义文本义原预测

杜家驹[1,2,3]，岂凡超[1,2,3]，孙茂松[1,2,3]，刘知远[1,2,3]

（1. 清华大学 计算机科学与技术系，北京 100084；
2. 清华大学 人工智能研究院，北京 100084；
3. 清华大学 智能技术与系统国家重点实验室，北京 100084）

摘　要：作为人类语言的最小语义单位，义原已被成功应用于许多自然语言处理任务。人工构造和更新义原知识库成本较大，因此义原预测被用来辅助义原标注。该文探索了利用定义文本为词语自动预测义原的方法。词语的各个义原通常都与定义文本中的不同词语的语义有相关关系，这种现象被称为局部语义相关性。与之对应，该文提出了义原相关池化(SCorP)模型，该模型能够利用局部语义相关性来预测义原。在 HowNet 上的评测结果表明，SCorP 取得了当前最好的义原预测性能。大量的定量分析进一步证明了 SCorP 模型能够正确地学习义原与定义文本之间的局部语义相关性。

关键词：义原预测；HowNet；语义相关性

中图分类号：TP391　　**文献标识码**：A

Lexical Sememe Prediction by Dictionary Definitions and Local Semantic Correspondence

DU Jiaju[1,2,3], QI Fanchao[1,2,3], SUN Maosong[1,2,3], LIU Zhiyuan[1,2,3]

(1. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
2. Institute of Artificial Intelligence, Tsinghua University, Beijing 100084, China;
3. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: Sememes, defined as the minimum semantic units of human languages in linguistics, have been proven useful in many NLP tasks. Since manual construction and update of sememe knowledge bases (KBs) are costly, the task of automatic sememe prediction has been used to assist sememe annotation. In this paper, we explore the method of applying dictionary definitions to predicting sememes for unannotated words. We find that sememes of each word are usually semantically related to different words in its dictionary definition, and we name this matching relationship local semantic correspondence. Accordingly, we propose a Sememe Correspondence Pooling (SCorP) model which is able to capture this kind of matching to predict sememes. Evaluated on HowNet, our model is revealed with state-of-the-art performance, capable of properly learning local semantic correspondence between sememes and words in dictionary definitions.

Keywords: sememe prediction; HowNet; semantic relevance

0　引言

在语言学中，义原(sememe)被定义为人类语言的最小语义单位[1]。一些语言学家认为，所有词语的意义都能够用一组数量有限的义原的组合来描述。由于我们不能直接从一个词语的形态特征中识别出其对应的义原，一些研究者预定义了一个义原集合并用其中的义原标注了许多词语，进而构成义原知识库(Sememe KB)。

HowNet[2]是目前最著名的义原知识库，其定义了大约 2 000 个语义无关的义原和约 90 种动态角色，并用这些义原和动态角色标注了超过 100 000 个中文和英文词语。HowNet 中的每个词语都含有

收稿日期：2019-10-24　**定稿日期**：2019-12-18
基金项目：国家自然科学基金(61661146007)

若干个义项，每个义项都用一棵由义原组成的树来表示，我们称之为义原树。图1展示了一个HowNet中词语义原标注的示例。词语“天文台”有一个义项，这个义项有4个义原，分别是“场所”“调查”“天象”“天体”，“agent”和“content”是义原之间的动态角色。HowNet已经被用于许多自然语言处理任务，例如，语义相似度计算[3]、词义消歧[4-5]、情感分析[6-8]、语言模型[9]、词表示学习[10]、词汇分类[11]等。

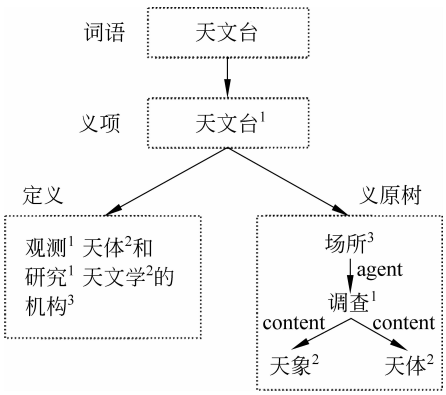

图1 词语“天文台”在HowNet中的义原标注和在词典中的定义

上标相同的义原和定义中的词语是语义相关的。

由于新词不断出现，已有词语的含义也在逐渐变化，但是人工为HowNet增加或更新义原标注费时费力。为了解决这一问题，Xie等[12]提出了自动义原预测任务，旨在为没有义原标注的词语自动推荐合适的义原。他们提出了两个简单有效的基于词向量的模型，这两个模型忽略了义原的层次结构，仅为词语预测一个义原的集合。Jin等[13]更进一步把词语中含有的汉字信息融合到义原预测模型中，提升了预测效果。

以上提到的方法都依赖于充分训练的词向量或字向量。而词向量或字向量的质量严重依赖于出现频率，因此这些方法很难为低频词和由低频字组成的词准确地预测义原。事实上，许多其他语言资源也能够用于义原预测。词典定义准确地解释了一个词语的含义，它非常容易获得，并且词典定义的质量与词频无关(也就是说低频词的定义质量也很高)。所以利用定义文本应当能够解决因低频词的词向量质量较差而引起的义原预测性能不佳的问题。Li等[14]最先提出在义原预测中使用定义文本，但是他们把从定义文本序列到义原的预测过程建模成一个序列到序列(Seq2Seq)问题，将并不恰当的顺序信息引入到词的义原集合之中。此外，他们还把定义文本看成字而非词语的序列——由于字的歧义性远大于词，这样会影响定义文本的编码质量。

本文提出了一个新的利用定义文本进行义原预测的模型。这一模型不仅解决了上述已有模型存在的问题，还能够刻画“局部语义相关性”。局部语义相关性是指义原树和定义文本之间的一种语义匹配关系。以图1为例，“天文台”的义原“场所”“调查”“天象，天体”分别在语义上与其定义文本中的词语“机构”“观测，研究”“天体，天文学”语义相关。这种相关性出现的原因是：一个词语的义原树和它的定义文本都描述了这个词语的含义。例如，在图1中，“天文台”的义原树可以被解释为“调查天象和天体的场所”，这与其定义文本非常相似。因为义原树可以被分解为若干个义原，定义文本能够分解为若干个词语，所以义原和定义文本的词语之间必定存在某种对应关系。

为了利用这种局部语义相关性，我们提出了义原相关池化(sememe correspondence pooling，SCorP)模型。SCorP模型首先计算了义原和定义文本中的词语之间的语义相关分数，然后在相关分数上做最大池化(max-pooling)操作。模型先预测义原的集合，然后通过预测义原之间的动态角色构造义原树。模型采用一个序列到集合的多标签分类框架来避免把不合适的顺序强加到义原集合上。此外，模型的输入是词而非字的序列，避免了汉字可能引入的歧义，同时还能够利用定义中词语已有的义原标注信息。最后，其还包含两种辅助操作，能够进一步提高义原预测的准确率。在实验中，我们评测了SCorP模型以及其他基线模型的义原预测性能，实验结果表明，SCorP模型取得了目前最好的结果。我们还做了一些量化分析，证明了SCorP模型能够正确地匹配义原和定义中的词语。

本文的贡献可以概括为：

(1) 发现了“局部语义相关性”，一种义原和定义文本中的词语之间的语义匹配关系。

(2) 提出了SCorP模型，能够利用局部语义相关性来预测义原，并取得了目前最好的义原预测结果。

1 相关工作

1.1 义原的应用

义原已经被广泛用于各项自然语言处理任务，

包括语义相似度计算[3]、词义消歧[4-5]、情感分析[6-8]等。此外，Niu 等[10]把义原信息加入到词表示学习中，利用义原来获取词语在上下文中的真实含义。Zeng 等[11]利用义原知识和注意力机制确定词语的类型，来自动扩充中文 LIWC 词表[15]。Gu 等[9]提出把义原当做语言知识专家，让它们在语言模型中帮助预测正确的词语。

1.2　义原预测

Xie 等[12]第一次提出了义原自动预测任务。他们还提出了基于协同过滤的义原预测模型 SPWE 和基于矩阵分解的方法 SPSE，取得了较好的义原预测效果。Jin 等[13]提出了两个能够利用词语含有的汉字信息的模型 SPWCF 和 SPCSE，以及一个同时考虑词语的外部信息和内部信息的集成模型 CSP。Li 等[14]第一次探索了定义文本在义原预测中的应用。他们提出了一个序列到序列模型 LD+Seq2Seq，其输入是词典中的定义文本或者 Wikipedia 中的描述文本的汉字序列，输出是预测出的义原序列。此外，还有一些工作尝试通过跨语言的义原预测来为其他语言构造义原知识库[16]。

1.3　定义文本的应用

在自然语言处理中，定义文本是一种重要且容易获得的语言资源。它们已经被用于各项自然语言处理任务，例如，词义消歧[17]、知识表示学习[18-19]、阅读理解[20]、反向词典[21-22]等。大多数之前的工作都把定义文本编码成一个向量作为下游任务的输入。另一些工作使用基于图的方法，在定义文本的基础上构造一个以词语为顶点的图[23-24]用于下游任务。

2　模型

在此部分，我们先引入一些必要的符号约定，然后简单介绍一个基础的序列到集合的多标签分类框架，这一框架是 SCorP 模型的基础。之后详细描述 SCorP 模型以及两种辅助操作。最后，简要介绍一个集成模型和义原树预测的方法。

2.1　符号约定

我们用“目标词语”来指代我们想要为之预测义原的词语。定义 W 为所有词语的集合，S 是所有义原的集合。给定一个词语 $w \in W$，其所有义原组成一个集合 $S_w = \{s_1, \cdots, s_{|S_w|}\}$，其中 $|\cdot|$ 表示集合的基数。词语 w 的定义表示为 $D_w = \{d_1, \cdots, d_{|D_w|}\}$，其中 d_i 是定义中的第 i 个词语。使用小写黑体符号表示向量，大写黑体符号表示矩阵，例如，$\boldsymbol{d}_i$ 是 d_i 的词向量，$\boldsymbol{s}_i$ 是 s_i 的义原向量，$\boldsymbol{D}_w$ 是由词向量 $\boldsymbol{d}_1, \cdots, \boldsymbol{d}_{|D_w|}$ 组成的矩阵。

2.2　序列到集合的多标签分类框架

序列到集合（sequence to set）的多标签分类（multi-label classification，MC）框架是我们提出的 SCorP 模型的基础，由两部分组成：一个编码器把定义文本编码成一个向量，一个多标签分类器用定义的向量计算每个义原的分数。最后选择分数高的义原作为模型的输出。

我们选择双向 LSTM（BiLSTM）[25]作为编码器。对于目标词语 w 的定义 $D_w = \{d_1, \cdots, d_{|D_w|}\}$，把定义中词语的预训练词向量 $\{\boldsymbol{d}_1, \cdots, \boldsymbol{d}_{|D_w|}\}$ 作为输入传递给 BiLSTM。然后 BiLSTM 输出两个隐状态序列，如式(1)所示。

$$(\overrightarrow{\boldsymbol{h}}_1, \cdots, \overrightarrow{\boldsymbol{h}}_{|D_w|}), (\overleftarrow{\boldsymbol{h}}_1, \cdots, \overleftarrow{\boldsymbol{h}}_{|D_w|}) = \mathrm{BiLSTM}(\boldsymbol{d}_1, \cdots, \boldsymbol{d}_{|D_w|}) \tag{1}$$

我们把两个方向上最后的隐状态拼接起来作为定义的向量 $\boldsymbol{v}$，然后把它作为输入传递给一个全连接层，如式(2)所示。

$$\begin{aligned} \boldsymbol{v} &= \mathrm{Concatenate}(\overrightarrow{\boldsymbol{h}}_{|D_w|}, \overleftarrow{\boldsymbol{h}}_1) \\ \boldsymbol{x} &= \boldsymbol{W}\boldsymbol{v} + \boldsymbol{b} \end{aligned} \tag{2}$$

其中，$\boldsymbol{W} \in \mathbb{R}^{|S| \times 2l}$，$\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^{|S|}$，$l$ 表示单个方向上的隐状态维度。$\boldsymbol{x}$ 的第 j 个元素 $[\boldsymbol{x}]_j$ 表示第 j 个义原的分数。在训练时，我们采用多标签一对多交叉熵损失函数（multi-label one-versus-all cross-entropy loss），如式(3)所示。

$$L = -\frac{1}{|S|}\sum_{j=1}^{|S|} [\boldsymbol{y}]_j \sigma([\boldsymbol{x}]_j) + (1 - [\boldsymbol{y}]_j)\sigma(-[\boldsymbol{x}]_j) \tag{3}$$

其中，$[\boldsymbol{y}]_j \in \{0, 1\}$ 表示第 j 个义原是否在词语 w 的义原集合中，σ 为 Sigmoid 函数。

2.3　SCorP 模型

MC 模型把整个定义编码为一个向量，这会导致信息损失，使其难以处理较长的定义。我们提出的义原相关池化（SCorP）模型利用了局部语义相关性，能够解决这一问题。

图 2 展示了一个用 SCorP 模型预测义原的例子。SCorP 模型的编码器与 MC 模型基本相同，但

是使用了式(1)中定义文本里每个词语的隐状态,而不是仅使用两个方向上最后的隐状态,如式(4)所示。

$$\boldsymbol{h}_i = \text{Concatenate}(\overrightarrow{\boldsymbol{h}_i}, \overleftarrow{\boldsymbol{h}_i})$$
$$1 \leqslant i \leqslant |D_w| \tag{4}$$

显然 $\boldsymbol{h}_i$ 包含了定义文本中第 i 个词语 d_i 的上下文信息。然后每一个 $\boldsymbol{h}_i$ 被作为输入传递给一个全连接层,输出组成一个矩阵,如式(5)所示。

$$\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}$$
$$\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{|D_w|}) \tag{5}$$

其中,$\boldsymbol{Y} \in \mathbb{R}^{|S| \times |D_w|}$ 是语义相关矩阵,$[\boldsymbol{Y}]_{ji}$ 表示第 j 个义原和定义文本中第 i 个词语之间的语义相关程度。由语义相关的局部性,词语的每一个义原都应该与定义中的一个或多个词语语义相关,与剩余的词语无关。也就是说,在义原预测中,一个义原是否出现在结果中取决于它是否与定义中的一些词语相关。因此,SCorP 模型通过在一个义原与定义中所有词语的语义相关程度上做最大池化(max-pooling)得到这个义原最终的分数,如式(6)所示。

$$[\boldsymbol{x}]_j = \max[\boldsymbol{y}_i]_j, \quad 1 \leqslant i \leqslant |D_w| \tag{6}$$

2.4 辅助操作

2.4.1 定义文本中词语的义原信息

我们把定义文本中词语的义原信息加入到编码器中。一方面,Niu 等[10]已经证实了义原信息有助于提升词向量的质量;另一方面,我们发现如果一个义原与一个定义中的词语语义相关,那么这个义原通常也是这个词语的一个义原。以词语"天文台"为例,它的一个义原"调查"与定义文本中的词语"观测"语义相关。同时,"调查"也是词语"观测"在 HowNet 中的一个义原。因此,定义文本中词语的义原信息应当有助于义原预测。

为了把定义文本中词语的义原信息融入到模型中,我们把义原向量的均值加到定义中的词语向量上,得到这个词语的新向量,如式(7)所示。

$$\boldsymbol{d'}_i = \boldsymbol{d}_i + \frac{1}{|S_{d_i}|}\sum_{s_j \in S_{d_i}} \boldsymbol{s}_j \tag{7}$$

其中,S_{d_i} 是 d_i 的义原集合,$\boldsymbol{d'}_i$ 是词语 d_i 的新向量。对于没有义原标注的词语,它的新向量与原来的向量相同。本文之后用"+SE"表示加入定义文本中词语的义原信息。

2.4.2 目标词语的向量

我们把目标词语的向量信息融入到模型中。如果我们已经获得了目标词语的词向量,就简单地把目标词语以及标点符号冒号(:)拼接在定义文本的前面。对于给定的词语 w,作为输入的词语序列为 $\{w, :, d_1, \cdots, d_{|D_w|}\}$。然后我们把这个新的词语序列转换为词向量序列,作为输入传递给 MC 或 SCorP 模型。

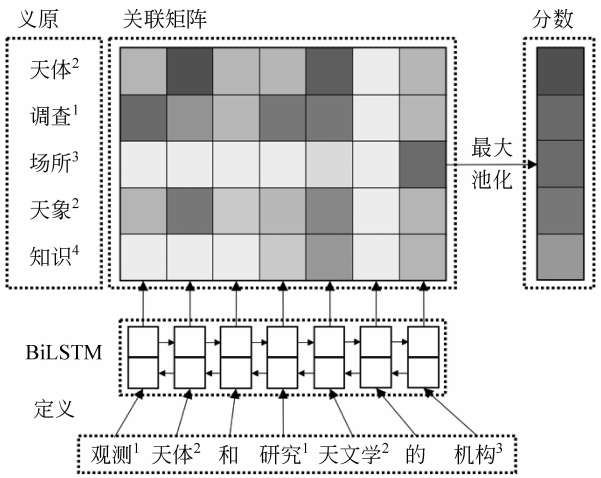

图 2 一个使用 SCorP 模型的例子示意图

在关联矩阵中,单元格的颜色越深,对应的义原和定义中词语的语义相关度越大。

为了能够处理没有词向量的词语,我们使用 BPE(byte pair encoding)[26]做了词语拆分。首先把没有词向量的词语拆分为更短的词语或汉字,使拆分出的子词都具有预训练的词向量。然后把拆分出的子词序列和标点符号拼接在词语的定义文本前面。本文之后用"+TW"表示考虑目标词语的向量。

2.5 集成模型

本文提出的模型主要使用词语的定义文本和词向量,而之前的模型 CSP[13]主要使用了目标词语的内部和外部信息。两种模型使用了不同的信息,我们把它们结合起来构造了一个集成模型。这个模型对每个义原的打分是两个子模型打分的加权平均。

2.6 义原树的预测

之前的义原预测工作都直接预测词语的义原集合,忽略了义原之间的动态角色。本文提出一种简单的从义原集合预测义原树的方法。在预测出义原集合 S 后,把 S 中任意两个义原 s_i, s_j 的向量表示与单词 w 的向量表示拼接起来,用一个多层神经网络预测这两个义原的动态角色,形成一个由义原作为

顶点、动态角色作为边的图。根据预测动态角色时的置信度，在图中删去置信度较低的边，直到图中没有环，就得到了这个单词的义原树。在训练时，把目标单词的义原树按照边拆分为“(义原，动态角色，义原)”这样的若干个三元组作为训练集。这一训练过程在之前提到的义原集合预测模型训练完毕后进行。

3　实验

3.1　数据集

我们使用了 HowNet 这一义原知识库。其标注了 103 843 个中文词语，103 390 个英文单词。这些词语和单词包含了 212 539 个语言无关的义项。在预测词语的义原集合时，我们过滤掉了所有在 HowNet 中出现次数少于 5 的义原，剩余的义原数量为 1 400。此外，我们把多义词的所有定义拼接起来得到一个较长的定义。

我们在中文和英文上评测了所有义原预测模型。在中文上，我们使用了《现代汉语词典》(第六版)①作为定义的来源。它定义了约 69 000 个汉字和词语的含义。我们用 THULAC[27] 把这些定义文本分割成词语。最终的数据集包含了 48 383 个词语，每个词语都在 HowNet 中有义原标注，在词典中有定义文本，并且还有对应的预训练词向量。我们用 Skip-gram[28] 方法在 Sogou-T② 语料库上预训练了词向量。数据集被随机地按照 8∶1∶1 的比例分割为训练集、验证集和测试集。

在英文上，我们把 WordNet[29] 作为定义的来源，使用 Glove[30] 预训练词向量，得到的数据集大小为 43 907，同样按照 8∶1∶1 的比例划分。中文和英文数据集的部分统计数据如表 1 所示。

表 1　数据集统计数据

统计项	中文	英文
词典大小	69 000	147 306
义项数量	98 732	206 978
HowNet 中的词语数量	103 843	103 390
有词向量的词语数量	5 606 977	400 000
数据集大小	48 383	43 907
平均义原数量	2.57	3.00

3.2　实验设置

3.2.1　基准方法

我们选择如下四种方法作为基准方法：

(1) SPWE[12]，一种基于词向量的方法，首先在词向量空间中找到与目标词相似的一些词语，推荐这些相似的词语的义原。

(2) CSP[13]，一个集成模型，由 4 个子模型组成，包含利用词语内部信息的 SPWCF 和 SPCSE 和利用词语外部信息的 SPWE 和 SPSE。

(3) LD＋Seq2Seq[14]，一个利用定义文本的序列到序列模型。

(4) MC，一种直接把定义文本编码成向量，然后进行预测的方法。

3.2.2　超参数选择与训练方法

所有模型的词向量维度均为 200，对于 SCorP 和 MC 这两个模型，BiLSTM 的隐状态维度为 256×2。对于其他模型，除词向量维度外，其他的超参数都进行了调整，直到在验证集上预测准确率达到最佳水平。对于集成模型，各个模型的权重为 λ_{SCorP} ∶ $\lambda_{\mathrm{CSP}}=1:10$。这个权重也是在验证集上调整得到的。

在训练时，我们使用 Adam 优化器[31]，学习率设置为 0.001。预训练的词向量在整个训练过程中是固定的。SCorP 和 MC 模型的各层都应用了 Dropout 操作。

3.2.3　评测方法

与之前的工作相同，我们使用 MAP(mean average precision)和 F_1 值作为评测标准。MAP 的计算方法为：假设一个词语有 K 个义原，模型根据打分从高到低对所有的义原排序，这 K 个义原在序列中的位置为 $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_K$，则 MAP 为如式(8)所示。

$$\mathrm{MAP}=\sum_{k=1}^{K}\frac{k}{p_k} \tag{8}$$

对于 F_1 值，除 LD＋Seq2Seq 外的所有模型都设置了一个阈值 δ。所有分数高于 δ 的义原组成模型的输出集合，用于计算 F_1 值。δ 也是一个超参数，同样在验证集上调整到了最佳值。我们采用了十折交叉验证，并报告十次实验的平均值。

① http://www.cp.com.cn/book/978-7-100-08467-3_44.html

② https://www.sogou.com/labs/resource/t.php

3.3 义原预测结果

表2列出了我们提出的所有模型和所有的基准模型在测试集上的义原预测结果。

表2 在测试集上的义原预测结果

模型	中文		英文	
	MAP	F_1	MAP	F_1
SPWE	55.04	48.23	40.56	38.25
CSP	58.93	50.26	42.58	37.56
LD+Seq2Seq	30.49	33.28	24.63	28.70
MC	51.24	42.06	49.09	39.71
MC(+TW)	59.15	48.77	54.56	45.34
MC(+SE)	53.99	45.90	52.62	45.24
MC(+TW,SE)	60.55	50.84	56.57	48.95
SCorP	54.95	46.89	56.17	50.41
SCorP(+TW)	63.46	53.07	59.70	52.89
SCorP(+SE)	57.57	49.99	58.28	52.83
SCorP(+TW,SE)	64.65	54.62	61.53	55.22
Ensemble	69.19	57.99	63.60	56.50

从表2可以得到：

(1) 模型SCorP(+TW,SE)取得了最好的义原预测效果，超过了之前的独立模型SPWE和LD+Seq2Seq，以及集成模型CSP。另外，我们提出的集成模型取得了目前最好的义原预测效果。

(2) SCorP模型的效果始终比对应的MC模型的效果好。这表明固定大小的向量不足以编码定义文本中的所有信息，考虑BiLSTM输出的所有隐状态是必要的。另外，最大池化操作能够有效地捕捉语义相关性。为验证这一点，我们比较了最大池化操作、平均池化操作(mean-pooling)和注意力机制(attention mechanism)这三种从BiLSTM输出的隐状态序列中采集信息的方法。平均池化是指对定义中每个词语对应的义原分数求均值。注意力机制是指对于每个义原，用义原的向量和BiLSTM输出的所有隐状态计算一个隐状态的加权平均作为上下文向量，这个向量经过一个全连接层得到这个义原的分数。我们省略了具体的计算过程，读者可以参考相关文章[32]。我们发现用平均池化或注意力机制替换最大池化后，MAP从54.95下降到了51.76和52.23。其原因是目标词语的义原通常仅与定义文本中的小部分词语语义相关，与剩余的其他词语完全没有关系。如果使用平均池化或注意力机制，一个义原的分数就会受定义文本中的全部词语的隐状态的影响，无关的词语会带来很多噪声。

(3) 在MC和SCorP模型的基础上，两种辅助操作(+SE,+TW)都能够提升义原预测的效果。+TW带来的提升较大，因为词向量能够有效地编码词语在上下文中的含义，之前的利用词向量的义原预测模型也取得了较好的效果。

(4) LD+Seq2Seq模型的效果最差，说明给义原规定一个顺序不利于义原预测。使用汉字的序列而不是使用词语的序列是另一个导致其效果变差的原因，因为汉字的多义性远大于词语的多义性，影响了对定义的编码。

3.4 未登录词的义原预测

依赖于目标词语的义原预测模型很难处理未登录词(out-of-vocabulary，OOV)，或者在处理OOV词语时性能严重下降。SCorP模型不需要词语的向量，因而能够处理OOV词语。+TW这一辅助操作也能够通过词语拆分处理OOV词语。

为了检测目前已有的模型在OOV词语上的效果，我们向数据集中添加了1 595个没有预训练词向量的词语，用新的数据集评测了所有模型。对于CSP模型，仅有两个子模型SPWCF和SPCSE能够预测OOV词语的义原，此时CSP模型变成了这两个模型的集成。

表3 未登录词的义原预测结果

模型	IV	OOV
CSP	59.03	42.92
LD+Seq2Seq	29.33	28.15
SCorP	54.44	51.05
SCorP(+TW)	63.10	54.30
SCorP(+TW,-WS)	62.97	51.20

注：IV(in-vocabulary)为已登录词，OOV(out-of-vocabulary)为未登录词。+TW,-WS代表在+TW操作中不进行词语切分。

表3展示了部分未登录词的义原预测结果。我们提出的SCorP模型及其变种SCorP(+TW)在OOV词语上的预测效果显著地好于两个基准模型。这说明我们提出的模型能有效地为新词预测义原。我们还通过实验验证了词语切分在处理OOV词语

时的作用。在 SCorP(＋TW)的基础上，不进行词语切分(＋TW，－WS)的模型在预测 OOV 词语的义原时的效果与不考虑目标词语信息的 SCorP 模型基本相同，说明词语切分在处理 OOV 词语时是必要的。

3.5　词语频率的影响

表 4 列出了四个模型在不同频率的词语上的义原预测效果。SCorP(＋TW，SE)模型在四类词语上的预测效果都是最佳的，稳定性最好。CSP 模型严重依赖于词向量，尽管在高频词上效果良好，但是在低频词(词频≤50)上预测效果下降十分严重。LD＋Seq2Seq 模型的效果基本不受词频影响，但是总体效果远差于其他模型。

表 4　不同频率的词语的义原预测结果

词语频数	≥5 000	500～5 000	50～500	≤50
CSP	62.18	59.60	46.60	26.34
LD＋Seq2Seq	29.31	31.83	33.61	31.93
MC(＋TW，SE)	63.18	61.75	59.08	54.32
SCorP(＋TW，SE)	65.42	64.71	62.94	58.91

注：四类词语的数量分别是 31 481、10 714、4 528 和 1 660。

3.6　案例分析

本节我们分析 SCorP 模型预测词语“天文台”的义原的全过程，说明 SCorP 模型能够有效地利用局部语义相关性。词语“天文台”含有 4 个义原“场所”“调查”“天象”“天体”。SCorP 模型能够输出所有词语和所有义原之间的语义相关度矩阵。表 5 展示了与每个词语最相关的 6 个义原及其分数。如果一个义原在某一行中为粗体，就表明这个义原和这一行中的词语的语义相关度最高。例如，在“机构”这一行，义原“场所”为粗体，表明在整个句子中，“机构”这个词语与“场所”的语义相关度最高。此外，“观测”与“调查”，“天体”与“天体”“天象”的语义相关度也是最高的。标点符号和虚词与所有义原的语义相关度都是比较低的。以上的观察表明 SCorP 模型能够有效地捕捉词语与义原之间的语义相关度。最后一行列出了在最大池化后分数最高的几个义原，4 个正确的义原恰好位于前 4 位。

表 5　词语“天文台”的义原预测结果分析

定义中的词语	预测出的义原
天文台	天体/0.20，告诉/－2.60，天象/－3.26，图像/－4.63，光/－4.65，知识/－5.00
：	天体/－4.37，告诉/－5.66，时间/－5.74，部件/－6.38，晨/－7.08，过去/－7.17
观测	**调查/2.37**，天象/－0.20，远/－1.28，天体/－1.35，测量/－2.19，看/－2.68
天体	**天体/6.16**，**天象/1.68**，调查/0.63，测量/－2.12，远/－3.40，看/－3.51
和	天体/－1.95，调查/－2.66，天象/－3.61，查/－4.26，看/－5.32，选择/－5.86
研究	调查/1.78，天象/－0.88，天体/－1.99，研究/－2.27，查/－2.62，知识/－3.46
天文学	天体/2.50，调查/1.54，天象/1.29，知识/0.56，白昼/－1.31，大地/－1.85
的	部件/－5.56，时间/－6.31，人/－6.36，地方/－7.04，告诉/－7.16，头/－7.74
机构	**场所/1.97**，天象/－1.32，知识/－1.48，天体/－2.02，设施/－2.14，用具/－2.90
。	时间/－1.66，天体/－3.02，部件/－3.21，调查/－4.90，天象/－4.95，白昼/－5.46
预测出的义原	**天体/6.16**，**调查/2.37**，**场所/1.97**，**天象/1.68**，知识/0.56，远/－1.28

注：第一列是这个词语在词典中的定义，第二列给出了定义中的每个词语在 SCorP 模型中预测出的义原及其分数。最后一行列出了模型在最大池化后预测的义原及其分数。

3.7　义原树预测结果

我们使用 edge-F_1 指标评测义原树预测的效果。edge-F_1 指标的计算方法是把真实的义原树和预测出的义原树按照边拆分成“(义原，动态角色，义原)”这样的三元组集合，然后计算真实的三元组集合和预测出的三元组集合之间的 F_1 值。

我们比较了 Nearest Neighbor 模型、MC 模型和 SCorP 模型的义原树预测的 edge-F_1 值，结果如表 6 所示，其中 Nearest Neighbor 模型是指根据词向量找到与目标词语最相似且已知义原树的词语，然后直接输出这个词语的义原树。我们提出的义原树预测方法的结果超过了 Nearest Neighbor 这一基

准模型。另外，SCorP 模型的义原树预测效果仍然高于 MC 模型，说明局部语义相关性在义原树预测中仍然有效。

表 6 义原树预测结果

模型	edge-F_1
Nearest Neighbor	43.45
MC(+TW,SE)	43.87
SCorP(+TW,SE)	44.46

4 结论

在本文中，我们发现了在利用定义的义原预测中的局部语义相关性，并以此为依据，提出了 SCorP 模型。我们还提出了两种辅助操作，包括加入定义文本中词语的义原信息和加入目标词语的词向量信息。实验结果表明，我们的模型取得了目前最好的预测效果。

我们之后还会在未来工作中探索以下研究方向：(1)定义中存在着主要成分和修饰成分，它们分别对应义原树中的顶层义原和其他义原，可能有助于义原树的预测。(2)把我们的模型应用到跨语言的义原预测中。

参考文献

[1] Bloomfield L. A set of postulates for the science of language[J]. International Journal of American Linguistics, 1926, 15(4):195-202.

[2] Dong Z, Dong Q. HowNet - A hybrid language and knowledge resource[C]//Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering, IEEE, 2003.

[3] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学，2002，7(2)：59-76.

[4] Zhang Y, Gong L, Wang Y. Chinese word sense disambiguation using HowNet[C]//Proceedings of International Conference on Natural Computation 2005. Springer, Berlin, Heidelberg, 2005：925-932.

[5] Duan X, Zhao J, Xu B. Word sense disambiguation through sememe labeling[C]//Proceedings of the 20th International Joint Conference on Arinfical in Telligence, 2007：1594-1599.

[6] 朱嫣岚，闵锦，周雅倩等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报，2006，20(1)：16-22.

[7] Xianghua F, Guo L, Yanyan G, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37：186-195.

[8] 党蕾，张蕾. 一种基于知网的中文句子情感倾向判别方法[J]. 计算机应用研究，2010，27(4):1370-1372.

[9] Gu Y, Yan J, Zhu H, et al. Language modeling with sparse product of sememe experts[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018：4642-4651.

[10] Niu Y, Xie R, Liu Z, et al. Improved word representation learning with sememes[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017(1)：2049-2058.

[11] Zeng X, Yang C, Tu C, et al. Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[12] Xie R, Yuan X, Liu Z, et al. Lexical sememe prediction via word embeddings and matrix factorization [C]//Proceedings of the 26th International Joint Conference on Artifical Ineelligence, 2017：4200-4206.

[13] Jin H, Zhu H, Liu Z, et al. Incorporating Chinese characters of words for lexical sememe prediction [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018：2439-2449.

[14] Li W, Ren X, Dai D, et al. Sememe prediction：Learning semantic knowledge from unstructured textual Wiki descriptions[J]. arXiv preprint arXiv：1808.05437, 2018.

[15] Pennebaker J W, Francis M E, Booth R J. Linguistic inquiry and word count：LIWC 2001[J]. Mahway：Lawrence Erlbaum Associates, 2001, 71：2001.

[16] Qi F, Lin Y, Sun M, et al. Cross-lingual Lexical sememe prediction[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018：358-368.

[17] Luo F, Liu T, Xia Q, et al. Incorporating glosses into neural word sense disambiguation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018：2473-2482.

[18] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proceedingsof the 30th AAAI Conference on Artificial Intelligence, 2016.

[19] Zhong H, Zhang J, Wang Z, et al. Aligning knowledge and text embeddings by entity descriptions[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015：267-272.

[20] Long T, Bengio E, Lowe R, et al. World knowledge for reading comprehension: Rare entity prediction with hierarchical LSTMs using external descriptions [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 825-834.

[21] Hill F, Cho K, Korhonen A, et al. Learning to understand phrases by embedding the dictionary [J]. Transactions of the Association for Computational Linguistics, 2016(4): 17-30.

[22] Bosc T, Vincent P. Auto-encoding dictionary definitions into consistent word embeddings[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 1522-1532.

[23] Thorat S, Choudhari V. Implementing a reverse dictionary, based on word definitions, using a Node-Graph Architecture [C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 2797-2806.

[24] Tissier J, Gravier C, Habrard A. Dict2vec: Learning word embeddings using lexical dictionaries[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 254-263.

[25] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[26] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1715-1725.

[27] Li Z, Sun M. Punctuation as implicit annotations for Chinese word segmentation[J]. Computational Linguistics, 2009, 35(4): 505-512.

[28] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of Advances in Neural Information Processing Systems,2013:3111-3119.

[29] Miller G A. WordNet: A lexical database for English [J]. Communications of the ACM, 1995, 38(11): 39-41.

[30] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[31] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412. 6980, 2014.

[32] Bahdanau Dzmitry, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409. 0473,2014.

杜家驹(1997—)，硕士研究生，主要研究领域为自然语言处理。
E-mail: djj18@mails.tsinghua.edu.cn

孙茂松(1962—)，通信作者，博士，教授，主要研究领域为自然语言处理。
E-mail: sms@tsinghua.edu.cn

岂凡超(1994—)，博士研究生，主要研究领域为自然语言处理。
E-mail: qfc17@mails.tsinghua.edu.cn