篇章级小句复合体结构自动分析

罗智勇* 北京语言大学 luo_zy@blcu.edu.cn
韩瑞昉 北京语言大学 15321103341@163.com
张明明 北京语言大学 15801572558@163.com
韩玉蛟 北京语言大学 chloe_hanyu@163.com
赵志琳 北京语言大学 Zhi_Lin_Zhao@163.com

摘要

话头话身共享关系是小句组合成小句复合体的重要语法手段，也是汉语篇章级句法语义分析的重要基础。本文通过引入窗口滑动机制，将篇章文本及其成分共享关系转换为文本片段及片段内部的成分共享关系预测问题，并针对预测结果合并与选择问题，依据话头话身共享关系的语法限定性，提出了多种候选项消除策略。实验结果表明，本文方法在缺少小句复合体边界信息条件下仍取得了与传统基于NTC的方法可比的实验结果，尤其是在确实缺失共享成分的待预测位置处的召回率提高了约0.4个百分点。

关键词： 小句复合体；边界；滑动窗口；预训练语言模型

Chinese Clause Complex Structure Automatic Analysis on Passage

Luo Zhiyong* 北京语言大学 luo_zy@blcu.edu.cn
Han Ruifang 北京语言大学 15321103341@163.com
Zhang Mingming 北京语言大学 15801572558@163.com
Han Yujiao 北京语言大学 chloe_hanyu@163.com
Zhao Zhilin 北京语言大学 Zhi_Lin_Zhao@163.com

Abstract

The naming-telling relationship is an important grammatical method for combining clauses into clause complexes, and it is also an important basis for Chinese text-level syntactic and semantic analysis. In this paper, by introducing the window sliding mechanism, the text and its component sharing relationship are transformed into the prediction problem of the text segment and the component sharing relationship within the segment. At the same time, in view of the problem of combining and selecting prediction results, this paper proposes a variety of candidate elimination strategies based on the grammatical limitation of the sharing relationship between words and phrases. The experimental results show that the method in this paper still achieves comparable experimental results with the traditional NTC-based methods in the absence of the boundary information of the clause complex, especially the recall rate at the to-be-predicted position where the shared components are indeed missing is improved by about 0.4 points.

Keywords: Clause complex , Boundary , Sliding window , Pre-Training language model

第二十一届中国计算语言学大会论文集，第623页-第633页，南昌，中国，2022年10月14日至16日。

1 引言

在中文自然语言处理中，通常会先对文本进行分句处理，再进行相关的下游任务。分句处理常见的做法是根据标点符号，如：句号、感叹号、问号、分号等，将文本分割成标点句序列。由于汉语标点句间存在成分共享的现象，这种机械分割方式会使得切分出来的标点句序列结构和意义不完整，从而直接影响阅读理解、机器翻译、信息抽取、搜索推荐等下游任务的性能。以下图1为机器翻译任务中的一个例子，包括中文原文及英语参考译文。

中文原文： 对于公民的申诉、控告或者检举，有关国家机关必须查清事实，负责处理。任何人不得压制和打击报复。
参考译文： The State organ concerned must, in a responsible manner and by ascertaining the facts, deal with the complaints, charges or exposures made by citizens. No one may suppress such complaints, charges and exposures or retaliate against the citizens making them.

Figure 1: 机械分割后标点句序列结构意义不完整示例

这段话引自中华人民共和国宪法第四十一条，中文原文由两个句号句组成，它的英语参考译文则引自全国人大网，同样也是两个句号句。其中，第二个句号句“任何人不得压制和打击报复”的受事是第一个句号句中的“公民的申诉、控告或者检举”。但它前面的句号隔断了这种关系，使“压制”和“打击报复”无法找到被施用者。如果机器翻译系统以句号句为单位进行翻译，则很难译出参考译文中加下划线的部分。由于句号句不一定能表示完整的意义，以句号句为单位进行语言信息处理的工作会受到本质性的影响。

针对标点句间成分共享现象，宋柔(2008)(2013)、尚英(2014)等人提出并完善了关于篇章层级的广义话题理论：小句复合体理论体系。小句复合体是语篇中的标点句序列，它是语篇的上下文语境中最大的紧密逻辑语义结构，也是最小的自足话头结构。自足话头结构即一个标点句序列既没有话头在上下文中，也没有词语可以看作上下文中某标点句的话头(宋柔, 2022)。

依据上述小句复合体的定义，界定小句复合体边界将依赖于标点句间的话头话身共享关系和逻辑语义关系。也就是说，在标点句间话头话身共享关系与逻辑语义关系分析清楚之前，很难界定小句复合体的边界。按照自然语言文本的认知方式，存在先有结构分析，后有边界的天然逻辑顺序。而目前关于小句复合体的结构自动分析的研究，都建立在人工标注好话头话身结构、划分好边界的小句复合体上，即先定好边界，而后进行结构分析。这样的做法颠倒了先结构分析、后有边界的逻辑顺序，不符合常规认知。

本文的研究决定回归人类对自然语言文本的认知方式，遵循先结构分析后边界的逻辑规律。即跳出小句复合体边界的限制，不再以边界完整的小句复合体为单位，而是将文本处理范围扩展到边界清晰的文本篇章上，进行基于滑动窗口的篇章级小句复合体结构自动分析。将问题转换为预测滑动窗口截取出的完整标点句序列上的每个标点句开头和结尾是否缺失成分，并找出缺失成分的开始位置和结束位置。再将所有窗口内部的结构拼接起来，形成整个篇章层级的小句复合体结构。

本文的主要贡献在于：（1）提出了在篇章长文本上进行结构自动分析的三步走策略：首先，将篇章文本及其成分共享关系转换为预训练语言模型可接受的样例级别的文本片段和片段内部的成分共享关系；其次，将样例经过预训练语言模型来学习和预测，得到每个样例对应的共享关系预测结果；最后，将样例级别的预测结果转换回篇章中去，得到篇章中每个标点句句首和句尾的共享成分在篇章中的位置。（2）提出了合并候选项的方法、是否重组答案、是否

基金项目：国家自然科学基金（62076037）

清除不合规预测答案共三种候选项消除策略，以及为缺失成分超出样例范围的待预测点设置不同标签类型的方法，来解决一个样例中包含多个待预测点时需要对多个预测结果进行合并的问题，以及待预测点缺失的成分不在样例文本范围内的问题。（3）在BERT-base、BERT-wwm、RoBERTa三个预训练语言模型上进行了实验，并与传统基于NTC的方法进行对比，对实验结果和预测错误的样例进行分析，验证了本文方法的有效性。

论文第2节介绍小句复合体结构自动的相关研究工作与进展；第3节介绍篇章级小句复合体自动分析问题的形式化定义；第4节介绍篇章级小句复合体模型的具体实现；论文第5节为实验验证与分析；第6节为论文总结与展望。

2 相关研究

宋柔指出，成分共享是构造小句复合体的基本语法手段。所谓成分共享(宋柔, 2022)，是指由于语言经济性需要，文本中话语片段中会有空缺成分，空缺的信息可能来自文本语境、话语交际场景、知识背景等。文本中某个话语片段字面上出现的一些成分，被另一些话语片段以字面上空缺的方式在语义上使用，这种现象就是成分共享。

汉语小句复合体理论体系中，包括四种成分共享模式(宋柔, 2022)，分别是话头共享模式（分支模式、新支模式、后置模式）、话身尾部共享模式（汇流模式）、超级小句复合体导引模式，以及模板模式。四种共享模式中，模板模式是在标点句中间的位置缺失成分，因此在现阶段的结构自动分析任务中难以形式化。其他三种成分共享模式都可以被形式化定义为话头话身识别任务。目前，已有许多研究者对这方面进行研究，主要可以分为基于传统机器学习的方法和基于深度学习的方法两类。

基于理论的传统方法主要由蒋玉茹等人提出。蒋玉茹(2012)首先根据话头分支模型，开展单个标点句的话头识别任务。其采用穷举策略，根据上一个话头自足句t_{i-1}为当前标点句c_i构造候选话头集合，然后利用编辑距离，计算候选话头和话头实例的相似性，筛选出正确的话头。在实现单个标点句的话头识别任务后，蒋玉茹(2017)又将话头识别任务扩展到标点句序列上。其仍采用穷举方法依次列举出每个标点句的候选话头，用树结构进行存储；再采用适当的策略计算话头候选树中每个节点的值，并计算每个叶节点到根节点的路径值；最后从中找到路径值最大的路径，从而得到标点句序列相对应的话头序列。但由于穷举策略会极大影响系统的执行效率，蒋玉茹(2014)又利用标点句在篇章中的位置特征、话头的语法特征、话头串和说明的邻接性等细粒度特征，指导候选话头的生成过程，尽可能的减少候选话头的个数，提高系统效率，在单个标点句和标点句序列上的话头识别任务的正确率均有提升。

基于深度学习的方法已有多人进行研究。M.Teng (2018)提出基于Attention-LSTM 的深度神经网络模型，进行单个标点句话头识别任务的研究，其实验结果相较于传统方法又有提升。但该研究在小句复合体话头共享关系自动分析方面，仅局限于分支模式，而对于新支模式、汇流模式、后置模式，以及多种模式混合的小句复合体结构的自动分析还未涉及。

针对多种成分共享模式，胡紫娟(2020)通过构建有向无环图NTCGraph结构来对小句复合体话头话身关系进行表示，并在预训练语言模型的基础上，通过Gather层分别取出待预测点向量和NTC向量，通过attention 交互层进行话头、话身的预测，进一步解决了之前只能够处理单一话头模式的问题。但该研究存在不缺失成分与缺失第一个字符冲突的问题，且处理的小句复合体长度小于128，不利于小句复合体结构自动分析的应用扩展。

因此，刘祥(2022)在胡紫娟的基础上开展进一步的研究。其首先引入空锚点机制，在输入的过程中通过插入特定的特殊token解决了因缺失成分位于句首处的标签与不缺失成分的标签重合问题；其次，针对特定插入待预测点的预测方式，提出了NT-MASK局部注意力机制，将全局注意力矩阵切分为了Sentence注意力矩阵和Mask-Sentence注意力矩阵，缓解了因插入待预测点对上下文信息编码带来的噪声干扰，减少不同待预测点之间的噪声干扰。除此之外，还构建了远距离共享成分识别数据集，将小句复合体输入长度由128扩充至512，进一步改善了汉语小句复合体自动分析的性能。

3 篇章级小句复合体自动分析形式化定义

篇章级小句复合体结构自动分析的任务主要可以分解为三个步骤。第一步，将篇章形式的文本处理为预训练语言模型可以接受的最大长度的文本样例；第二步，对样例文本内部的标点句句首和句尾的成分共享关系进行识别；第三步，将每个样例预测出的结果合并为篇章的结

果。本节将对第一步和第三步中小句复合体话头话身关系形式化定义、转换和还原，预测结果合并和选择，以及篇章级小句复合体结构自动分析问题描述进行介绍，图2为主要流程。

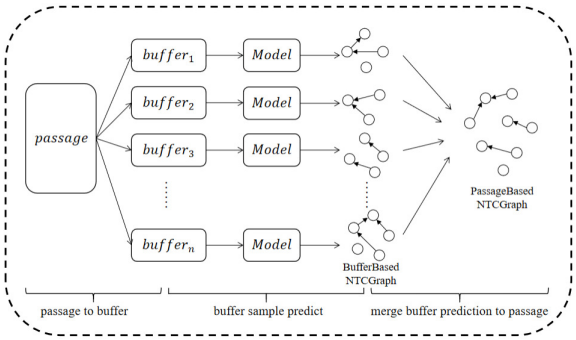

Figure 2: 篇章级小句复合体结构自动分析流程

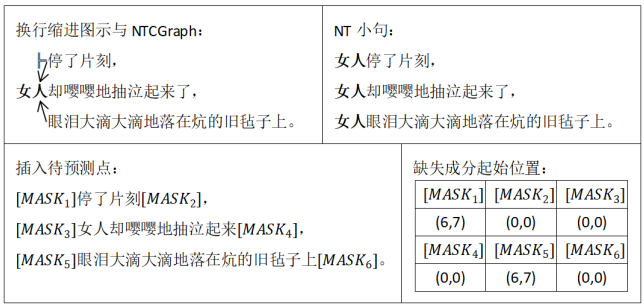

Figure 3: 成分共享关系形式化定义图示

3.1 成分共享关系形式化定义

胡紫娟首次将小句复合体的话头话身结构定义为广义有向无环图NTCGraph(胡紫娟, 2020)，表示为G=(V,E)。V是结点的集合，一个结点表示NTC中的一个标点句。E是小句复合体中的每两个标点句间话头话身成分共享关系的集合。NTCGraph与小句复合体一一对应。篇章级小句复合体话头话身成分共享关系则由一个篇章中多个NTCGraph线性相连而来，表示为Passage-NTCGraph，包含了篇章中每个标点句句首句尾缺失的共享成分位置。如图3所示，该NTC中包含三个标点句，换行缩进图示与NTCGraph中的箭头由标点句句首或句尾指向其所缺失的共享成分，NT小句为补充共享成分后的句子。在每个标点句的句首和句尾各插入一个[MASK]作为待预测点，每个MASK处缺失的共享成分位置则为在文本中的开始和结束位置。

3.2 成分共享关系转换与还原

由于预训练语言模型对输入长度的限制，成分共享关系需要由篇章级转换为预训练语言模型可接受的样例级，以方便预训练语言模型的学习和预测。每个待预测点的成分共享关系转换公式如下：

$$(start_b, end_b) = \begin{cases} (0,0), & start_p = 0, end_p = 0 \\ (0,0), & start_p < buf_s \, or \, end_p > buf_e \\ (start_p - buf_s, end_p - buf_s), & start_p > buf_s \, and \, end_p < buf_e \end{cases}$$

$(start_b, end_b)$表示每个待预测点的共享成分在样例中的相对位置；$(start_p, end_p)$是从Passage-NTCGraph中获取的该待预测点的共享成分在篇章中的绝对位置；buf_s和buf_e分别指样例中第一个字符与最后一个字符在篇章中的位置。

样例经过模型预测之后，再将生成的样例级预测结果转换为篇章级。即将预测得到的样例级成分共享关系还原回篇章级的成分共享关系。每个待预测点的成分共享关系还原公式如下：

$$(start_p, end_p) = \begin{cases} (0,0), & start_b = 0, end_b = 0 \\ (start_b + buf_s, 0), & start_b \neq 0, end_b = 0 \\ (0, end_b + buf_s), & start_b = 0, end_b \neq 0 \\ (start_b + buf_s, end_b + buf_s), & start_b \neq 0, end_b \neq 0 \end{cases}$$

3.3 预测结果合并与选择

由于一个待预测点会出现在多个样例中，对应地会产生多个$(start_b, end_b)$，这些预测结果位置在各自的样例文本中不同，但还原回篇章中的$(start_p, end_p)$后可能相同，因此需要对相同的预测结果进行合并，最终从不同的预测结果中选取一个作为该待预测点处的成分共享位置。

本文将还原后的每一个篇章中的$(start_p, end_p)$及其产生概率作为一个候选项，加入到对应待预测点的所有预测结果中。对于相同的$(start_p, end_p)$，将其对应的产生概率进行取平均、求和、或取最大的运算来进行合并，合并后的概率作为该$(start_p, end_p)$的最终概率。根据最终概率最大的原则从多个候选项中选择最佳的作为该待预测点处的最终预测结果。合并的过程如图4所示。

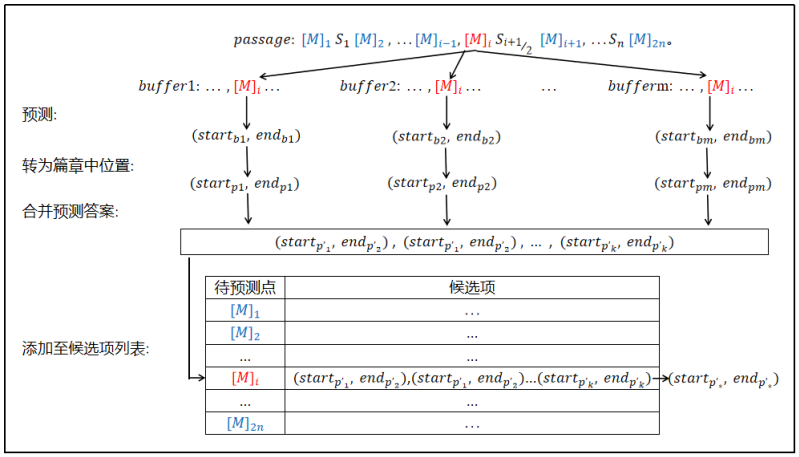

Figure 4: 预测结果合并与选择流程

3.4 篇章级小句复合体结构自动分析问题描述

成分共享关系的转换和还原定义之后，篇章级小句复合体结构自动分析转化为在样例窗口内部进行话头话身关系识别的任务。任务定义为：对于每个标点句的开始和结束位置，在给定的样例文本范围内预测每个字符位置可能成为该待预测位置缺失成分的开始位置和结束位置的概率，也就是预测样例窗口内每个标点句的开头和结尾是否缺失成分，并找出缺失成分在窗口中的开始位置和结束位置。输入输出格式定义如下：

输入的格式为：在标点句两端插入标志[MASK]，形成待预测的信息点，开头插入[CLS]，结尾插入[SEP]。输出的格式则为每个[MASK]位置所缺失的成分在窗口的文本（未插入[MASK]等符号)中开始和结束的索引位置(start,end)。如果不缺失成分，则为(0,0)。图3提到的NTC对应的输入输出形式如图5所示。样例内部的成分共享关系识别即定义为，对于样例中的每个待预测点，在给定的输入文本片段中选择该待预测点缺失成分的开始位置和结束位置，与抽取式机器阅读理解任务的答案形式(Cui et al., 2019b)相同。

Input:

[CLS][MASK]停了片刻[MASK],[MASK]女人却嘤嘤地抽泣起来[MASK],[MASK]眼泪大滴大滴地落在炕的旧毡子上[MASK]。[SEP]

Output:

(6,7), (0,0), (0,0), (0,0), (6,7), (0,0)

Figure 5: 输入输出格式

4 模型方法

4.1 模型架构

小句复合体结构自动分析模型大体仍采用胡紫娟(2020)和刘祥(2022)基于NTC的方法中使用的框架。模型共包括五层：输入层、编码层、收集层、注意力层和输出层。图6是模型的整体框架。

输入层的Embedding表示由Token Embeddings、Segment Embeddings、Position Embeddings三者相加而来。本文的Segment Embeddings不需要区分句子，因此统一用'0'来表示。Position Embeddings则是每个token 的位置向量表示。输入的最终表示为：

$$E = (E_{[CLS]}, E_{m1}, E_{t1}, E_{t2}, E_{m2}, ..., E_{tn}, E_{mm}, E_{[SEP]})$$

编码层的作用是对文本的上下文信息进行编码。该层基于BERT预训练语言模型，将每个token的embedding表示，经过12层Transformer Encoder Blocks来获得结合了上下文语义信息的语义表示，每一层的输出作为下一层的输入，并将Transformer最后一层的输出作为编码层的输出。

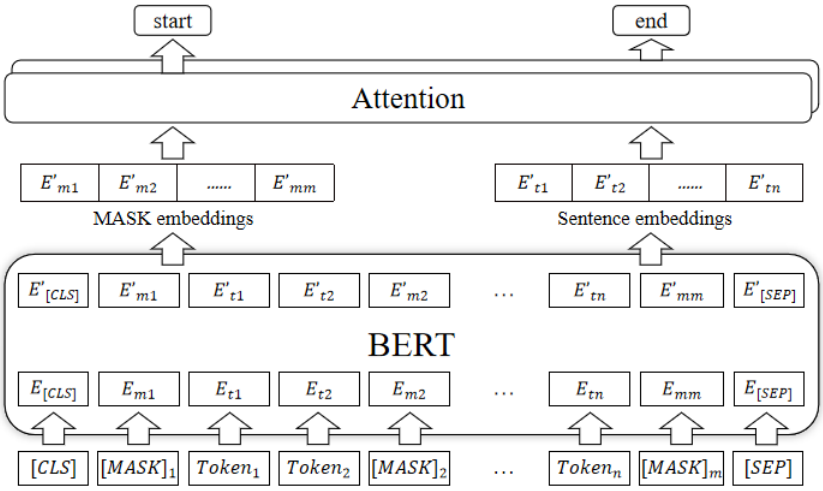

Figure 6: 窗口内标点句序列话头话身关系识别模型整体架构

收集层的作用是，按照在处理语料时记录的[MASK]与非[MASK]的位置，将样例中所有的待预测点[MASK]和文本信息分别抽取出来，得到MASK_embeddings矩阵和Sentence_embeddings矩阵。

注意力层的核心任务是充分结合文本信息，对MASK_embeddings和Sentence_embeddings的相似性进行计算，最终获得每个MASK在每个文本信息字符上能够作为答案的开始位置和结束位置的概率。该层主要依据多头注意力机制(Vaswani et al., 2017)，将上一层的输出MASK_embeddings和Sentence_embeddings 分别看作Quary和Key进行匹配。Attention分为start_attention和end_attention。前者负责预测缺失成分的起点，后者负责预测缺失成分的终点。这样即可在一个[MASK]位置获取两个可以代表start_logits和end_logits的结果。

输出层为最后一层，主要任务是将所有[MASK]对应的在每个文本token处作为答案的开始和结束的概率转化成答案的开始和结束位置，期望输出格式为(start,end)，用于找到其对应的话头话身。start和end结果的维度为[batch_size,MASK_length,1]，即将start_logits中一个[MASK]对应的多个概率值，转化为一个[MASK]对应一个位置索引。

实现的方法是将上一层得到的start_logits和end_logits经过softmax来得到两个使得成为答案的概率值最高的位置索引。计算公式如下：

$$P_s(p_m = i|M) = softmax(H_m^s \cdot H_t^{sT})[i]$$

$$P_e(p_m = i|M) = softmax(H_m^e \cdot H_t^{eT})[i]$$

前者表示MASK处预测的start位置为文本中第i个token的概率，后者表示MASK处预测的end位置为文本中第i个token的概率。$H_m^s, H_t^s, H_m^e, H_t^e$上标的s和e分别表示是对开始和结束位置的预测，下标m和t分别表示MASK和非MASK的token。模型通过端到端的过程来学习。Loss的计算分为对[MASK]处缺失的共享成分开始位置start的计算，和对共享成分的结束位置end的计算，并将二者的平均值作为模型整体的loss表示。计算公式如下：

$$L_s = -\sum_{m=1}^{n} \log P_s(p_m|M)$$

$$L_e = -\sum_{m=1}^{n} \log P_e(p_m|M)$$

$$Loss = \frac{L_s + L_e}{2}$$

4.2 候选项消除策略

基于滑动窗口的方法中，同一待预测点会出现在多个窗口样例中，对应地产生多个预测结果，需要将多个窗口预测出的结果（窗口中的相对位置）转换为篇章中的结果（绝对位置）。每个预测结果及其产生概率(start,end,prob)作为该待预测点的一个候选项，问题转换为如何在多个候选项中选出最佳作为最终预测结果。本文从候选项合并方法、是否重组答案，以及是否清除不合规预测答案三个方面来考虑。

候选项合并方法即3.3节中对相同$(start_p, end_p)$产生概率的合并方法，包括概率求平均、求和、求最大。同时还试验了不合并候选项的方法，以及去除同一个待预测点下为零候选项的方法。去除为零候选项指当候选项中有不为(0,0)的候选项时（即预测该位置缺失成分），去除掉为(0,0)的候选项，最终结果从不为(0,0)的候选项中选择概率最大者；否则不去除。

重组答案是指将预测得到的答案对拆分开来进行开始和结束位置的重组的方式。即将同一个待预测点得到的全部开始位置和结束位置重新进行组合。在拆分重组的过程中，开始和结束位置应当遵循答案的合法性原则，即如果开始位置为0，只能和结束位置为0组合，以及答案的开始位置应当小于或等于答案的结束位置等。

清除不合规预测答案是指清除掉模型预测得到的不合法的答案。不合规分为以下几类：答案对中有一个位置为0而另一个位置不为0；答案的开始位置大于结束位置；答案中的共享成分跨越了一个以上标点句；答案中的成对标点符号不匹配……将不合规候选答案从候选项列表中清除之后，有利于减少错误答案的干扰，选出正确答案。

4.3 缺失成分不在样例内部的处理策略

基于滑动窗口的方法使用滑动窗口截取文本中的标点句序列，这样会截断窗口内部和外部的成分共享关系，导致如果窗口内成分不完整的待预测点的共享话头话身在窗口范围外，则无法正确预测到其缺失的内容。

对于这种待预测点，本文采取了以下几种处理策略：第一，赋予(0,0)的标签，即在样例文本内其缺失成分的位置指向[CLS]；第二，赋予其样例末尾最后一个字符的位置的标签，即在样例文本内其缺失成分的位置指向[SEP]的位置；第三，在第一种策略的基础上增加第三位标签来表示该待预测点在篇章中的成分是否完整，‘0’表示该待预测点在篇章中成分完整，不缺失成分；‘1’表示该待预测点在篇章中成分不完整，缺失成分；第四，为每个待预测点单独构造一条样例，并确保该待预测点缺失的成分处于样例文本内部，来避免这个问题。以上策略中除第一种方式和不缺失成分的待预测点标签相重合之外，其他三种方式均将这类缺失成分不在样例内部的待预测点和不缺失成分的待预测点的标签区别开来。

5 实验

5.1 数据集

本文的语料为北京语言大学中文小句复合体标注语料（CBBC）(宋柔, 2017)，包含百科、报告、新闻、小说共4个领域。经过对语料的重新清洗后，共计9种类别，245个篇章，12509个NTC，40379个标点句，80758个待预测点。

为保证训练集、验证集和测试集之间没有重合的篇章，在每种类别下按7:2:1的比例划分篇章，得到训练集、验证集和测试集。按照滑动窗口数据构造格式，以及NTC为单位的数据格式，分割后的数据集样例个数如下表1所示：

数据集样例个数	train	eval	test	总计
滑动窗口样例	21240	4101	6845	32186
NTC样例	8155	1818	2534	12507

Table 1: 不同格式的数据集样例个数

5.2 评估指标

本文采取的评估指标包括篇章中缺失成分的待预测点处的精确率、召回率、F1值，以及篇章中全部待预测点的总正确率。具体评估指标如下：

精确率Precision: 缺失成分的待预测点预测正确的个数占预测结果为缺失成分的待预测点总个数的比重（当且仅当开始位置和结束位置均正确时，该待预测点正确）。定义为：$Precision = N_{lost}/N_{predict_{lost}}$。

召回率Recall: 缺失成分的待预测点预测正确的个数占目标答案中缺失成分的待预测点总个数的比例。定义为：$Recall = N_{lost}/N_{target_{lost}}$。

F1值：精确率和召回率的调和平均。定义为：$F1 = (2 * Precison * Recall)/(Precison + Recall)$。

总正确率Accuracy：预测正确的待预测点个数（包括不缺失成分的待预测点和缺失成分的待预测点）占所有待预测点总个数的比例。定义为：$Acc_{total} = (N_{lost} + N_{nlost})/(N_{total_{lost}} + N_{total_{nlost}})$。

5.3 实验结果

本小节使用基于滑动窗口的数据集在BERT-base(Devlin et al., 2018)，BERT-wwm(Cui et al., 2019a)，RoBERTa(Liu et al., 2019)三个预训练语言模型上进行了实验。实验以概率求和合并候选项作为候选项合并方法，并清除不合规候选项。基线方法为刘祥(2021)的传统的限定小句复合体边界的样例构造方式，即利用人工标注的小句复合体边界信息，以边界完整的小句复合体为单位进行训练和测试。二者最终得到的每个样例的预测结果都对应回篇章中的绝对位置，且用相同测试篇章进行评测。下表2为实验结果：

	Precision	Recall	F1	Acc_total
Baseline	0.7345	0.7490	0.7417	0.9103
BERT-base	0.6638	0.7293	0.6950	0.8867
BERT-wwm	0.6622	0.7432	0.7003	0.8859
RoBERTa	0.6842	**0.7530**	0.7169	0.8923

Table 2: 不同预训练语言模型上的实验结果

实验表明，基于滑动窗口的数据集在RoBERTa上，缺失成分的待预测点处精确率、召回率、F1值都达到最佳，分别为0.6842的精确率、0.7530的召回率、0.7169的F1值，在全部待预测点（包括缺失成分的待预测点和不缺失成分的待预测点）处的正确率达到0.8923。其次为BERT-wwm，BERT-base上效果最差。

与baseline基于NTC的方法相比，基于滑动窗口的方法在缺失成分的待预测点处召回率达到历史最佳水平，比baseline基于NTC的方法的召回率0.7490提高0.4个百分点，这体现了基于滑动窗口的方法对缺失成分的识别能力有所提升。而在其他三个指标下，基于滑动窗口的方法表现稍逊于基于NTC的方法，但也在一定程度上达到了可比的性能。这是由于基于滑动窗口的方法难度更大。

第一，基于NTC的方法使用独立的NTC作为输入，没有像基于滑动窗口方法那样带来NTC之外的干扰信息。滑动窗口方法的输入包含了大量的标点句文本信息，文本长度最大程度接近512，模型需要从近512个字符中去学习和预测两个位置，包含了大量的冗余信息。而NTC方法输入文本长度为NTC长度，待预测点缺失成分一定在NTC范围内，不包含NTC之外的干扰信息。

第二，基于NTC的方法每个待预测点只产生一个预测结果，无需对预测结果进行筛选合并。基于滑动窗口的方法同一待预测点出现在多个样例中，相应地会产生多个预测结果。虽然预测结果中会有NTC方法得不到的正确结果，但如何从多个预测结果中筛选出正确答案仍是难点。

第三，基于NTC的方法每个待预测点在样例中的标签具有一致性，而基于滑动窗口的方法同一个待预测点在不同样例中的标签不一致。后者同一待预测点在不同样例中位置不同，答案标签也不同。更有共享成分超出文本范围的情况，使得同一个待预测点的标签既有真正缺失成分的答案位置，又有(0,0)或[SEP]的位置。标签不一致给模型的训练和预测造成一定的困难。

因此，基于滑动窗口的方法与基于NTC的方法相比难度更大，基于NTC的方法占优是在情理之中。

5.4 候选项选择策略的研究

基于滑动窗口的方法使得同一待预测点处有多个窗口产生的多个预测结果候选项，针对这一问题，本文从合并候选项的方法、是否重组答案、是否清除不合规预测答案三个方面来考虑。不同的候选项合并方法下缺失成分的待预测点处P、R、F1值如表3所示:

	合并候选项		不合并候选项
	avg-prob	sum-prob	
不去零项	0.7265/0.2573/0.3800	0.7186/0.6959/0.7071	0.8105/0.4409/0.5712
去除零项	0.6788/0.7435/0.7097	0.6886/0.7442/0.7153	0.6832/0.7495/0.7148
	重组答案对		
	avg-prob	sum-prob	max-prob
不去零项	0.7119/0.2648/0.3860	0.7176/0.6978/0.7075	0.8036/0.4547/0.5808
去除零项	0.6733/0.7447/0.7072	0.6774/0.7492/0.7115	0.6767/0.7485/0.7108
	清除不合规答案		
	avg-prob	sum-prob	max-prob
不去零项	0.7327/0.2578/0.3813	0.7250/0.6994/0.7120	0.8147/0.4412/0.5724
去除零项	0.6803/0.7487/0.7129	0.6842/0.7530/0.7169	0.6839/0.7528/0.7167

Table 3: 不同候选项选择策略在测试集上的P、R、F1值

表中1-4行表示采取不同合并候选项方法且不重组答案对时，缺失成分待预测点的P、R、F1值。包括从候选答案中去除零项、将相同开始结束位置候选项的概率值合并、概率求平均、求和、求最大等情况。5-8行表示重组答案对与合并候选项相结合的实验结果。9-12行表示清除不合规答案且不重组答案对时，与合并候选项相结合的实验结果。

实验结果表明，不重组答案对、清除不合规答案，且与概率求和的候选项合并方法结合的情况下，缺失成分的待预测点处的召回率、F1值达到最佳表现。尤其是召回率，达到了0.7530。这是由于概率求和的方法下，某个(start,end)出现的次数越多，最终概率值就越大。即预测到某个(start,end)的窗口个数越多，该候选项就越可能成为正确答案。同时，重组答案对与不重组的方法相比，为无正确答案对的待预测点增加了正确的答案对候选项，有助于将正确答案从候选项列表中被筛选出来，提升了缺失成分待预测点的召回率。而清除不合规答案后，将模型预测得到的不合规的答案对从候选项列表中剔除掉，降低了这类概率高但错误的候选答案对正确答案的影响。也就是在缺失成分的待预测点处，如果候选项中包含正确答案，清除不合规答案的方法有助于将正确答案从候选项列表中被筛选出来，作为最终答案。

5.5 缺失成分不在样例内部的处理策略研究

本小节对4.3节中提到对于缺失成分不在样例内部的四种处理策略进行了实验，实验结果如表4所示。其中CLS_Labels表示第一种指向[CLS]的位置，也就是5.3中Roberta上的实验结果；SEP_Labels表示第二种指向[SEP]的位置；Triple_Labels表示第三种三位标签的策略；Single_Sample表示第四种样例中只包含单个待预测点的策略。

	Precision	Recall	F1	Acc_total
CLS_Labels	0.6842	**0.7530**	0.7169	0.8923
SEP_Labels	0.6754	0.7503	0.7108	0.8902
Triple_Labels	0.7088	0.7335	**0.7209**	**0.8959**
Single_Sample	**0.7171**	**0.6769**	0.6964	0.8878

Table 4: 缺失成分不在样例内部的不同处理策略实验结果

这几种处理策略中，CLS_Labels在缺失成分的待预测点处召回率达到历史最佳水平，为0.7530。其只利用了每个待预测点处缺失成分的位置信息，在对多个预测结果合并时采取了候选项概率求和合并以及清除不合规候选答案的措施，使得真正缺失成分的待预测点的召回率最高。SEP_Labels将缺失成分不在样例内部的待预测点与不缺失成分的待预测点用标签区

分开来，帮助模型更好地识别缺失成分与不缺失成分的待预测点，再加上候选项选择策略，达到了稍弱于CLS_Labels的性能。Triple_Labels的实验将每个待预测点在篇章中是否成分完整的信息融合进来，使得最终的答案预测和多个结果合并时候选答案的选取更加合理。其F1值和Acc_{total}最高，综合结果在三者中表现最好。Single_Sample的实验则直接避免了多个预测结果合并这一难点，其实验结果中对于缺失成分的待预测点处Precision在三种方法中最高。

5.6 错误样例分析

从上述实验结果来看，基于滑动窗口的方式综合实验结果很难超越以NTC为样例的实验结果。前一小节中我们对实验结果进行了宏观的对比和分析，本节中对测试集的15262个待预测点中两种方法预测错误的样例进行详细分析。

基于滑动窗口的方法预测错误的待预测点共1643个（缺失成分987个,不缺失成分656个）。基于NTC的方法预测错误的待预测点共1374个（缺失成分1008个，不缺失成分366个）。虽然滑动窗口的方法预测错误的总数大于NTC的方法，但对于缺失成分的待预测点，滑动窗口的方法预测错误的数量小于NTC的方法，即滑动窗口方法对于缺失成分的待预测点上的话头话身识别更加准确。同时，基于滑动窗口预测错误的987个缺失成分待预测点中，有197个（近20%）待预测点的候选答案中包含正确答案，只是其产生概率较低，无法在答案选择环节被选为最终答案，但这仍然表明了滑动窗口方法的潜力所在。

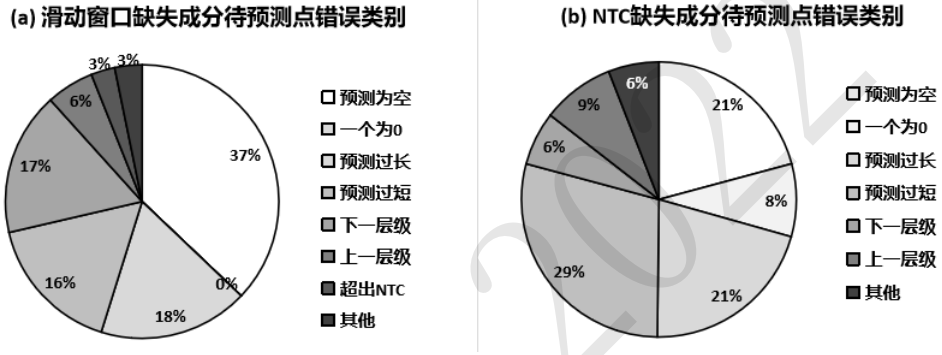

Figure 7: 缺失成分待预测点的错误类型

除此之外，对于缺失成分的待预测点来说，从答案长度、答案的层级、超出NTC范围来对比NTC方法和滑动窗口方法。各错误类型及占比如图7所示。NTC方法错误的类型为预测过长和过短的比例为50%，滑动窗口方法这一项的比例为34%，这说明NTC方法更容易预测为一个答案位置正确，而另一个答案位置错误，使得超出正确答案范围或得到正确答案的一部分，而滑动窗口方法预测一个答案位置正确、另一个答案位置错误的比例较低。NTC方法错误的类型为上一层级和下一层级的比例为15%，滑动窗口方法这一项的比例为23%，说明了滑动窗口方法在层级的学习上表现不如NTC的方法，错误地预测为上一层级或下一层级的情况更多。除此之外，滑动窗口预测错误的一个重要类型是超出NTC范围，这也是滑动窗口方法的难点之一，即滑动窗口的输入范围比NTC方法的输入范围更大，使得预测得到的结果会超出NTC文本。

6 总结与展望

基于NTC的方法存在很大的限制，那就是需要明确的NTC边界信息，这在测试场景中很难满足。基于滑动窗口的方法突破了这种限制，虽然综合结果并未超过NTC方法，但在缺失成分的待预测点处的召回率取得最优，其他指标结果也仍具有可比性，这表明基于滑动窗口的方法客观上是有意义和一定潜力的。除此之外，如何将汉语小句复合体结构自动分析应用到其他下游任务中去，也至关重要。何晓文(2021)等人提出基于小句复合体的句子边界自动识别研究，王瑞琦(2021)和刘祥(2021)等人提出将小句复合体结构自动分析模型与阅读理解任务相结合，在一定程度上提升了阅读理解任务的性能。

未来的工作将对现有的方法进行总结和反思，从当前错误样例总结的规律中继续展开研究，进一步提升小句复合体结构自动分析任务的性能，并探索小句复合体结构自动分析的应用。

参考文献

Y. Cui, W. Che, T. Liu, B. Qin, Z. Yang, S. Wang, and G. Hu. 2019a. Pre-training with whole word masking for chinese bert.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China, November. Association for Computational Linguistics.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Y. Jiang and R. Song. 2017. topic structure identification of pclause sequence based on generalized topic theory *.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Xiang Liu, Ruifang Han, Shuxin Li, Yujiao Han, Mingming Zhang, Zhilin Zhao, and Zhiyong Luo. 2021. Shared component cross punctuation clauses recognition in chinese. In Lu Wang, Yansong Feng, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 709–720, Cham. Springer International Publishing.

M. Teng, Y. Zhang, Y. Jiang, and Y. Zhang. 2018. Research on construction method of chinese nt clause based on attention-lstm. In *Ccf International Conference on Natural Language Processing & Chinese Computing*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv*.

Ruiqi Wang, Zhiyong Luo, Xiang Liu, Rui Han, and Shuxin Li. 2021. 基于小句复合体的中文机器阅读理解研究(machine reading comprehension based on clause complex). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 723–735, Huhhot, China, August. Chinese Information Processing Society of China.

何晓文, 罗智勇, 胡紫娟, and 王瑞琦. 2021. 基于小句复合体的句子边界自动识别研究. 中文信息学报, 35(5):8.

刘祥. 2022. 汉语小句复合体结构自动分析改进策略研究. Master's thesis, 北京语言大学.

宋柔. 2008. 现代汉语跨标点句句法关系的性质研究. 世界汉语教学, (2):19.

宋柔. 2013. 汉语篇章广义话题结构的流水模型. 中国语文, (6):12.

宋柔. 2017. 小句复合体的理论研究和应用. https://2011.gdufs.edu.cn/info/1070/2085.htm.

宋柔. 2022. 小句复合体的语法结构.

尚英. 2014. 汉语篇章广义话题结构理论的实证性研究. Ph.D. thesis, 北京语言大学.

胡紫娟. 2020. 汉语小句复合体话头结构分析. Master's thesis, 北京语言大学.

蒋玉茹and 宋柔. 2012. 基于广义话题理论的话题句识别. 中文信息学报, 26(5):114–120.

蒋玉茹and 宋柔. 2014. 基于细粒度特征的话题句识别方法. 计算机应用, 34(5):5.