文章编号：1003-0077（2011）00-0000-00

基于小句复合体的句子边界自动识别研究*

何晓文，罗智勇，胡紫娟，王瑞琦

（北京语言大学 信息科学学院，北京 100083）

摘要：自然语言文本的语法结构层次是词、短语、句子、小句复合体、语篇。词、短语相关处理技术已经相当成熟，而句子的概念至今未有公认的适用于语言信息处理的界定。语言学界对于句子的定义缺乏操作性；而且以句号句为单位进行语言信息处理的工作会受到本质的影响，这种状况阻碍了汉语信息处理的发展。小句复合体是对于逻辑语义关系最小自足的小句序列的复合结构。本文把汉语的句子大致地界定为自足话题结构。基于小句复合体的句子边界识别对句法语义分析、机器翻译等具有一定的指导意义。本文基于小句复合体理论，采用 BERT 边界识别模型对句子的边界进行识别。实验结果表明，该模型对句子边界自动识别精确率、召回率、F1 值分别达到 91.91%、88.87%、90.36%，该模型对句子边界识别效果远好于按照不同的标点符号机械分割的效果。

关键词：句子；小句复合体；句子边界识别

中图分类号：TP391　　**文献标识码：**A

Automatic Recognition of Sentence Boundary Based on Clause Complex

HE Xiaowen, LUO Zhiyong, HU Zijuan, WANG Ruiqi

(School of Computer Science, Beijing Language and Culture University, Beijing, 100083, China)

Abstract: The grammatical structure of natural language texts consists of word, phrase, sentence, clause complex and text. The processing technology of word and phrase has been quite mature, but the concept of sentence has not yet been recognized as applicable to the processing of language information. The definition of sentence in linguistic circles is lack of operability; moreover, the processing of language information by using full-stop sentence as a unit will be essentially affected, which hinder the development of Chinese information processing. The clause complex is a composite structure of clause sequences which minimal self-sufficient in logical and semantic relations. In this paper, Chinese sentences are roughly defined as self-contained topic structures. Sentence boundary recognition based on clause complex has certain guiding significance for syntactic and semantic analysis, machine translation and so on. Based on the theory of clause complex, this paper uses the BERT boundary recognition model to recognize the boundary of sentence. The experimental results show that the accuracy, recall rate and F1 value of the model are 91.91%, 88.87% and 90.36% respectively, the recognition effect of the model is much better than that of mechanical segmentation according to different punctuation marks.

Key words: sentence; clause complex; sentence boundary recognition

0. 引言

篇章处理的基本单位是句子。由于句子是很多诸如句法语义分析、机器翻译、语音识别、智能问答等自然语言处理应用系统的输入，句子边界切分错误，将在一定程度上降低相应应用系统的性能。句子边界问题主要包括以下两个方面。

* **收稿日期：2019.6.18　　定稿日期：2019.7.31**

基金项目：北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（19YCX124）；北京市哲学社会科学规划研究基地项目（１３ＪＤＺＨＢ００５）

作者简介：

何晓文（1994——），女，硕士，主要研究方向为自然语言处理；

罗智勇（1975——），通讯作者，男，副教授，硕士研究生导师，主要研究方向为语言信息处理、机器学习；

胡紫娟（1993——），女，硕士，主要研究方向为自然语言处理；

王瑞琦（1995——），女，硕士，主要研究方向为自然语言处理；

（1）语言学定义中，句子的定义缺乏操作性；一是缺少语言学规范的约束。真实文本中句号的使用常常带有一定的随意性，因此以句号切分的句号句不具备当作基本语法单位的资格。二是结构和意义不完整。一般人的印象中句号句是结构和意义完整的，但事实并非如此。

例 1：安军在长安杀安禄山仇视的政敌及其家属；对投降的官僚则迁到洛阳，授以官爵。又大肆搜括民财，弄得民间骚然不安。（百科全书）

c1:安军在长安杀安禄山仇视的政敌及其家属；

c2: 对投降的官僚则迁到洛阳，

c3: 授以官爵。

c4: 又大肆搜括民财，

c5: 弄得民间骚然不安。

图 1 含有 5 个标点句的换行缩进图示

例 1 中的百科全书语料包含 5 个标点句，分别为 c1、c2、c3、c4、c5。以句号切分例 1 中的语料，得到 c1c2c3、c4c5 两部分内容，从例 1 的换行缩进图示图 1 中我们观察到：c1c2c3 标点句序列的内容表达的意思完整，但 c4c5 标点句序列的内容表达的意思并非完整，标点句 c4、c5 都缺少主体安军，安军位于标点句 c1 中。因此，句号不能作为分割句子的标记。

（2）实际自然语言处理应用中，往往直接以句号、分号、问号、叹号等标点符号作为句子的分割；但实际上句号等并不能作为句子的分割标记。我们建立了一个包含新闻、政府工作报告、小说、百科全书等多领域文本构成的话头话体共享关系标记语料库（以下简称**小句复合体语料库**），百科全书分库共 4757 个句号句，其中 2202 个共享前面句号句中的成分，占 46.3%，比例很高。同样规模的政府工作报告分库和小说分库中这类情况分别占 14.9%和 5.6%，比例虽不很高但也不能无视。下面以机器翻译应用为例说明。

例 2：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。（中华人民共和国宪法第四十一条）[1]

例 2 由汉语的两个句号句组成，它的英语参考译文（引自全国人大网）也是两个句号句：

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.

例 2 中，“任何人不得压制和打击报复。”的受事是“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。

本文把汉语的句子大致地界定为自足话题结构[2]。句子对于各种汉语文本具有认知意义、支持篇章处理的各种应用等性质。使得句子的切分成为重要的研究课题。

本文以标点句序列作为输入，将句子边界识别问题转化成机器学习有指导分类问题，并构建了基于 BERT[3]的边界识别模型，对汉语句子边界进行识别。实验结果表明本文提出的句子边界识别模型效果远好于按照句号等标点符号机械分割结果。

本文第 1 节介绍相关研究；第 2 节介绍相关概念并对句子的边界识别进行建模；第 3 节介绍 BERT 边界识别模型；第 4 节介绍实验，第 5 节为总结与展望。

1. 相关研究

虽然已有大量研究者关注汉语篇章分析[4,5]，但实际上着重句子分析的研究工作并不多，对于怎样界定汉语的句子，至今还缺少深入的研究。

赵元任[6]提出：句子是最大的语法分析上重要的语言单位。一个句子是两头被停顿限定的一截话语。这种停顿应理解为说话的人有意做出的。其中“最大”缺少可操作的检验标准。朱德熙[7]提出：句子是前后都有停顿并且带有一定的句调表示相对完整的意义的语言形式。其中“停顿”和“句调”是语音标志，在文本中是部分地可检验的。“相对完整的意义”则缺少可操作性。左思民[8]对汉语句子的构成和定义进行分析，他提出：句子是按照句子模式，通过言语活动而构成。从本质上下定义，句子是一个交际功能相对自足的言语单位，是根据小句模式生成的言语单位。宋柔[2]把汉语的句子大致地界定为自足话题结构。之所以说“大致地”，是因为有时一个自足话题结构因带有某些连词而逻辑上不能独立，需要与和它相邻的作为逻辑关联方的自足话题结构合在一起，才能构成汉语的句子。

谷晶晶[9]提出了一种基于句子的分词与词性标注信息进行汉语逗号自动分类的方法。核心工作是特征的筛选与抽取。他们分别采用最大熵模型和 CRF 模型[10]构建逗号分类器，实现对汉语逗号的七元分类与识别任务。谷晶晶又提出了一种针对汉语冒号的标注体系与识别方法。他们分别采用使用规则法和最大熵模型实现冒号的自动分类与识别。目前自然语言处理领域针对小句复合体的句子边界自动识别比较少，使得小句复合体的句子边界自动识别成为重要的研究课题。

本文基于 BERT 边界识别模型对句子的边界进行识别。实验结果表明，该方法对句子的边界识别精确率、召回率、F 值分别达到了 91.91%、88.87%、90.36%，该模型对句子的边界识别效果远好于按照不同形式的标点符号机械分割的效果，该研究对句法语义分析、机器翻译等输入单元问题具有一定的指导意义。

2. 一些概念与小句复合体的句子边界识别建模

2.1 标点句

我们把逗号、分号、句号、问号、叹号、直接引语的引号以及这种引号前面的冒号所分隔出的词语串称为标点句[11]。

2.2 小句复合体

小句复合体是话头共享关系、逻辑语义关系和指代关系都不可分割的最小的标点句序列。作为汉语文本的信息处理单位，汉语篇章中小句复合体的结构关系至少包括 3 个方面：话头话体关系，逻辑语义关系，指代关系。话头话体关系用以在小句复合体内拆出小句，而逻辑语义关系是小句间的关系，指代关系的某些所指也是小句。

本文目前的工作只涉及小句复合体的话头共享关系。

2.3 句子

如果一个广义话题结构既没有话题在上下文中，也没有说明在上下文中，它就称为自足的话题结构[2, 13,14]，简称自足话题结构，也就是汉语的句子。某个标点句如果不是语境中其他标点句中话题的说明，且没有成分做话题被语境中其他标点句说明，这种标点句称为话题结构独立句，简称独立句，独立句也是句子。

例 3: １８７５年（光绪元年），清政府采纳左宗棠的建议，派军进入新疆。左宗棠采取≮先北后南，缓进速战≯的进军方针，于１８７６年收复天山北路，１８７７年进入南疆。阿古柏在节节失败、众叛亲离情况下于库尔勒服毒自杀，其汗国亦随之覆灭。（百科全书）

①c1:１８７５年（光绪元年），

c2:　　清政府采纳左宗棠的建议，

c3:　　派军进入新疆。

②c4:左宗棠采取≮先北后南，缓进速战≯的进军方针，

c5:　　于１８７６年收复天山北路，

c6:　　１８７７年进入南疆。

③c7:阿古柏在节节失败、众叛亲离情况下于库尔勒服毒自杀，

c8:其汗国亦随之覆灭。

图 2 含有 8 个标点句的小句复合体结构

例 3 中包含 8 个标点句，分别为 c1、c2、c3、c4、c5、c6、c7、c8，组成 3 个小句复合体①、②、③，小句复合体间以空行分隔。①中包含一个句子，为 c1c2c3；②中包含一个句子，为 c4c5c6；③中包含 2 个句子，分别为 c7、c8。标点句 c1 缺少说明**清政府采纳左宗棠的建议**，标点句 c2 缺少话题１８７５年（光绪元年），标点句 c3 缺少话题１８７５年（光绪元年）和清政府。标点句 c1、c2、c3 组成的整体既没有话题在上下文中，也没有说明在上下文中，故标点句 c1、c2、c3 组成的整体是一个完整的句子。标点句 c4 成分完整，但是被标点句 c5 共享左宗棠，标点句 c6 缺少话题左宗棠和于，标点句 c4、c5、c6 是一个句子。标点句 c7、c8 都为独立句，故标点句 c7、c8 都是句子。

2.4 机器学习问题的描述

本文将基于小句复合体的句子边界识别问题转换为标点句序列的二分类问题。该二分类模型可以定义为：给定一个标点句对 X = <x1，x2>，x1、x2 为标点句序列，若在标点句序列 x1 为前一个小句复合体的结尾、而 x2 是下一个小句复合体的开始，则 X 对应的标记 Y 为 1，其它情况 Y 为 0。

汉语的句子涉及多个标点句，我们对例 4 中的语料进行标注。分别给出 X 标点句序列个数为 2、3 的两种标注形式。

语料的换行缩进形式如图 3 所示，标点句序列个数为 2 的语料的标注信息如图 4 所示。标点句序列个数为 3 的语料的标注信息如图 5 所示。

例 4：１９９５年末居民储蓄存款余额接近 3 万亿元，比≮七五≯末增加两万多亿元。城乡劳动就业不断增加。脱贫工作取得很大成绩，贫困人口由≮七五≯末的８５００万减少到 6

５００万。(政府工作报告)

c1: １９９５年末居民储蓄存款余额接近3万亿元，

c2: 比<七五>末增加两万多亿元。

c3: 城乡劳动就业不断增加。

c4: 脱贫工作取得很大成绩，

c5: 贫困人口由<七五>末的８５００万减少到６５００万。

图3 含有7个标点句的小句复合体结构

0 １９９５年末居民储蓄存款余额接近3万亿元， 比<七五>末增加两万多亿元。

1 比<七五>末增加两万多亿元。 城乡劳动就业不断增加。

1 城乡劳动就业不断增加。 脱贫工作取得很大成绩，

1 脱贫工作取得很大成绩， 贫困人口由<七五>末的８５００万减少到６５００万。

图4 标点句个数为2的语料的标注形式

1 １９９５年末居民储蓄存款余额接近3万亿元，比<七五>末增加两万多亿元。 城乡劳动就业不断增加。

1 比<七五>末增加两万多亿元。城乡劳动就业不断增加。 脱贫工作取得很大成绩，

1 城乡劳动就业不断增加。脱贫工作取得很大成绩， 贫困人口由<七五>末的８５００万减少到６５００万。

图5 标点句个数为3的语料的标注形式

3. BERT边界识别模型

BERT（Bidirectional Encoder Representations from Transformers）是基于双向Transformer的大规模预训练语言模型，该语言模型能高效抽取文本中的信息并应用于各种NLP任务。图6为BERT边界识别模型。

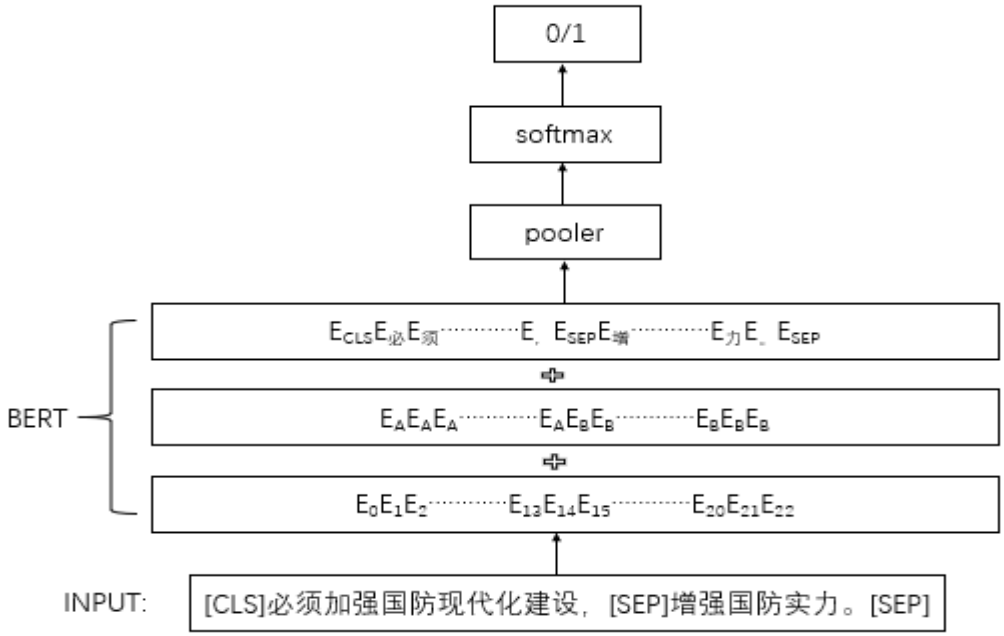

图 6 BERT 边界识别模型

BERT 边界识别模型主要由以下几个网络层构成：第一层为输入；第二层为 BERT 预训练模型；第三层为 pooler 层；第四层为 softmax 层。其中 BERT 预训练模型的向量由三部分词向量拼接得到，分别为词向量(Token Embedding)、位置向量(Position Embedding)、句向量(Segment Embedding)。其中位置向量表示该词的位置信息，NLP 中单词顺序是特别重要的特征，需要对位置信息进行编码，句向量用于表示输入的标点句对信息。将 BERT 模型在[CLS]位置处的输出经过 pooler 层，全连接层和非线性激活层，此时的输出向量既包括词的信息，又包括上下文语义的信息。最后通过 softmax 层对小句复合体标点句序列进行分类。BERT 分类模型总的前向计算公式如下所示：

$$y = \mathrm{softmax}_{\theta 1}\left(pooler_{output\ \theta 2}\left(BERT_{\theta 3}\left(Content_{left}, Content_{right}\right)\right)\right)$$

$$\theta(\theta 1,\ \theta 2,\ \theta 3)$$

为了避免过拟合问题，我们在损失函数中加入了刻画模型复杂程度的指标 L2 正则化。L2 正则化的计算公式为：

$$R(\theta) = ||\theta||_2^2 = \sum_i |\theta_i^2|$$

4. 实验

4.1 数据集

本文采用北京语言大学中文小句复合体标注语料，包括百科全书、政府工作报告、新闻、小说 4 个领域的语料，共有 12675 个句子。语料中句子的分布情况如下：

表 1 句子的分布情况

语料类别	句子的个数
百科全书	4471

政府工作报告	4017
新闻	1798
小说	2389
总计	12675

我们依据小句复合体理论对百科全书、政府工作报告、新闻、小说语料进行标注。语料中共标出了 37090 个标点句对，我们将按照 7：2：1 的比例划分训练集、验证集、测试集。

4.2 机械分割语料的识别结果

在标注好的语料的基础上，我们分别按照句号、句号分号、句号叹号、句号问号 4 种形式的标点符号对百科全书、政府工作报告、小说、新闻文本中的语料进行机械分割。其中按句号、句号分号、句号叹号、句号问号正确分割的小句复合体的句子的标点句对的正确率分别为：81.44%、80.08%、81.65%、81.60%。结果如下表所示。

表 2 是按句号分割语料的分布情况。全部语料中分割正确的标点句对的正确率为 81.44%，其中政府工作报告、百科全书、新闻、小说语料中分割正确的标点句对的正确率分别为 83.08%、79.65%、82.78%、81.64%。

表 2 按句号分割语料的分布情况

语料	标点句对数	分割正确的标点句对数	正确率
政府工作报告	10093	8385	83.08%
百科全书	13751	10953	79.65%
新闻	4819	3989	82.78%
小说	8427	6880	81.64%
总数	37090	30207	81.44%

表 3 是按句号分号分割语料的分布情况。全部语料中分割正确的标点句对的正确率为 80.08%，其中政府工作报告、百科全书、新闻、小说语料中分割正确的标点句对的正确率分别为 82.17%、77.00%、82.28%、81.37%。

表 3 按句号分号分割语料的分布情况

语料	标点句对数	分割正确的标点句对数	正确率
政府工作报告	10093	8293	82.17%
百科全书	13751	10588	77.00%
新闻	4819	3965	82.28%
小说	8427	6857	81.37%
总数	37090	29703	80.08%

表 4 是按句号叹号分割语料的分布情况。全部语料中分割正确的标点句对的正确率为 81.65%，其中政府工作报告、百科全书、新闻、小说语料中分割正确的标点句对的正确率分别为 83.55%、79.63%、83.05%、81.88%。

表 4 按句号叹号分割语料的分布情况

语料	标点句对数	分割正确的标点句对数	正确率
政府工作报告	10093	8433	83.55%
百科全书	13751	10950	79.63%
新闻	4819	4002	83.05%
小说	8427	6900	81.88%
总数	37090	30285	81.65%

表 5 是按句号问号分割语料的分布情况。全部语料中分割正确的标点句对的正确率为 81.60%，其中政府工作报告、百科全书、新闻、小说语料中分割正确的标点句对的正确率

分别为 83.08%、79.64%、83.75%、81.82%。

表 5 按句号问号分割语料的分布情况

语料	标点句对数	分割正确的标点句对数	正确率
政府工作报告	10093	8385	83.08%
百科全书	13751	10951	79.64%
新闻	4819	4036	83.75%
小说	8427	6895	81.82%
总数	37090	30267	81.60%

4.3 实验设置

模型参数：实验中所有模型使用深度学习开源框架 Tensorflow 搭建。训练集的 Batch size 设置为 32，验证集的 Batch size 设置为 32。优化器的学习率为 0.00005。迭代次数为 2000。

4.4 实验结果

该实验对句子的边界进行识别，标点句序列个数为 2 的实验在测试集上的精确率、召回率、F1 值分别为 90.24%、89.69%、89.96%。由上文我们得到按句号、句号分号、句号叹号、句号问号机械分割的标点句对的正确率分别为：81.44%、80.08%、81.65%、81.60%，正确率分别提高了 8.80%、10.16%、8.59%、8.64%。表 6 为标点句序列个数为 2 的句子边界自动识别结果。

表 6 标点句序列个数为 2 的句子边界自动识别结果

语料	精确率	召回率	F1 值
政府工作报告	89.09%	85.37%	87.19%
百科全书	90.50%	90.30%	90.40%
新闻	87.70%	91.15%	89.39%
小说	95.63%	95.48%	95.55%
全部语料	90.24%	89.69%	89.96%

标点句序列个数为 3 的实验在测试集上的精确率、召回率、F1 值分别为 91.91%、88.87%、90.36%。由上文我们得到按句号、句号分号、句号叹号、句号问号机械分割的标点句对的正确率分别为：81.44%、80.08%、81.65%、81.60%，正确率分别提高了 10.47%、11.83%、10.26%、10.31%。表 7 为标点句序列个数为 3 的句子边界自动识别结果

表 7 标点句序列个数为 3 的句子边界自动识别结果

语料	精确率	召回率	F1 值
政府工作报告	94.04%	89.87%	91. 91%
百科全书	90.53%	89.79%	90.16%
新闻	88.24%	91.94%	90.05%
小说	94.22%	95.64%	94.93%
全部语料	91.91%	88.87%	90.36%

我们对政府工作报告、百科全书、新闻、小说领域的语料分别进行研究，对比不同领域句子的边界识别效果。表 8 为不同领域语料实验对比结果。

一、通过对比实验我们得到 BERT 边界识别模型对句子的边界自动识别的效果好于按照不同形式的标点句机械切分的效果。

二、实验表明：通过 BERT 边界识别模型对句子边界自动识别进行研究，标点句序列个数为 3 的句子边界识别效果好于标点句序列个数为 2 的句子边界识别效果。

三、对政府工作报告、百科全书、新闻、小说 4 个领域的语料进行研究，实验表明：相比于政府工作报告、新闻语料，句子的边界自动识别效果在百科全书、小说语料上效果好。

表 8 不同领域语料实验对比结果

语料	按句号分割语料的正确率	按句号分号分割语料的正确率	按句号叹号分割语料的正确率	按句号问号分割语料的正确率	标点句序列个数为 2 的句子边界自动识别结果	标点句序列个数为 3 的句子边界自动识别结果
政府工作报告	83.08%	82.17%	83.55%	83.08%	89.09%	94.04%
百科全书	79.65%	77.00%	79.63%	79.64%	90.50%	90.53%
新闻	82.78%	82.28%	83.05%	83.75%	87.70%	88.24%
小说	81.64%	81.37%	81.88%	81.82%	95.63%	94.22%
全部语料	81.44%	80.08%	81.65%	81.60%	90.24%	91.91%

4.5 实验结果分析

例 5：我们要为经济振兴打好基础，必须使基本建设保持必要的规模，但这个规模一定要同国力相适应，不能超过财力负担和物资供应的可能。如果违反这一客观经济规律，就会受到现实生活的惩罚。（**政府工作报告**）

c1:我们要为经济振兴打好基础，
c2:　　必须使基本建设保持必要的规模，
c3:但这个规模一定要同国力相适应，
c4:　　不能超过财力负担和物资供应的可能。
c5:如果违反这一客观经济规律，
c6:　　就会受到现实生活的惩罚。

图 7 含有 6 个标点句的小句复合体结构

我们以例 5 为例对句子边界自动识别结果进行分析。例 5 中包含 6 个标点句，分别为 c1、c2、c3、c4、c5、c6，组成一个小句复合体。小句复合体中包含 c1c2、c3c4、c5c6 三个句子。BERT 边界识别模型能够正确的识别出标点句 c1c2、c2c3、c3c4、c5c6 之间的关系，但未能正确的识别出标点句 c4c5 之间的关系。标点句 c4、c5 之间存在转折关系。

对于语料中没有正确识别出句子边界的标点句序列，我们发现部分标点句序列中存在转折、对比、假设等关系。

针对发现的问题，我们将不断地改进模型，并尝试通过 BERT-CRF、BERT-BLSTM-CRF 两种方法分别对句子的边界进行识别。

5. 总结与展望

小句复合体是语篇的基本语义单位，其句内、外的功能组织在语篇的发展过程中扮演了重要的语义角色。句子之间关系是小句复合体关系中基础性的内容。本文基于小句复合体理论对句子的边界进行识别，通过实验结果我们得到在政府工作报告、百科全书、新闻、小说语料上 BERT 边界识别模型对句子的边界自动识别效果好于按照不同形式的标点句机械切分的效果。通过 BERT 边界识别模型对句子边界自动识别进行研究，标点句序列个数为 3 的句子边界识别效果好于标点句序列个数为 2 的句子边界识别效果。其中，句子边界自动识别效果在百科全书、小说上的效果好于政府工作报告、新闻语料上的效果。

下一步的工作中，由于我们目前的工作标注的标点句对只涉及相邻的两个标点句和三个标点句，汉语的句子涉及的不仅仅是相邻两个标点句和相邻的三个标点句，所以我们将不断的扩大标注语料标点序列的规模。其次，不断地改进模型，并尝试通过 BERT-CRF、BERT-BLSTM-CRF 两种方法分别对句子的边界进行识别。最后，小句复合体中含有嵌套的

成分，我们目前的工作还没有提出区分嵌套的小句复合体的句子边界的方法。例如：

例 5：他还认为，就汽车营销而言，中国现在还几乎是一张白纸，这意味着难得的发展机遇。中国汽车工业的进步将强有力地推动中国经济的发展。

c1:他还认为，

c2: 【就汽车营销而言，

c3: 中国现在还几乎是一张白纸，

c4: 这意味着难得的发展机遇。

c5: 中国汽车工业的进步将强有力地推动中国经济的发展。】

图 6 含有 5 个标点句的小句复合体结构

目前我们还未对例 5 中嵌套的成分**就汽车营销而言，中国现在还几乎是一张白纸，这意味着难得的发展机遇。中国汽车工业的进步将强有力地推动中国经济的发展**进行分析，也没有提出区分句子边界的方法。扩大标点序列的规模、如何从嵌套结构中识别出句子的边界是下一步的研究工作。

参考文献

[1] Shili Ge & RouSong, The Naming Sharing Structure and its Cognitive Meaning in Chinese and English, SedMT2016(Workshop of NAACL2016), San Diego, USA, June 16, 2016.

[2] 宋柔，葛诗利，尚英等. 面向文本信息处理的汉语句子和小句. 中文信息学报, 2017(3) .

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding.arXiv:1810.04805.

[4] Emily Pitler, A.Louis and A.Nenkova. Automatic Sense Prediction for Implicit Discourse Relations in Text.[C].ACL2009:683-691.

[5] Emily Pitler and A. Nenkova. Using Syntax to Disambiguate Explicit Discourse Connectives in Text.[C].ACL2009:13-16.

[6] 赵元任. 汉语口语语法[M].吕叔湘译.北京：商务印书馆，1979.

[7] 朱德熙. 语法讲义[M].北京：商务印书馆，1982.

[8] 左思民. 汉语句子的构成和定义.上海师范大学学报（哲学社会科学版）,1988(1):1004-8634 .

[9] 谷晶晶. 汉语逗号与冒号的自动分类识别研究[D].苏州大学.7714.

[10] J Lafferty, A McCallum, F Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of ICML,2001:282-289.

[11] 宋柔.汉语篇章广义话题结构的流水模型[J].《中国语文》,2013(6):483-494.

[12] 宋柔.小句复合体的理论研究和应用.[DB/OL].http://2011.gdufs.edu.cn/info/1070/2085.htm,2017-11-13.

[13] 尚英.汉语篇章广义话题结构理论的实证性研究[D].北京：北京语言大学.2014.

[14] 卢达威，宋柔，尚英.从广义话题结构考察汉语篇章话题认知复杂度[J].中文信息学报.2014,28(5):112-124.

作者联系方式：
何晓文，北京语言大学信息科学学院，100083，18810098695，2907506523@qq.com
罗智勇，通讯作者，北京语言大学信息科学学院，100083，18201105723，luo_zy@blcu.edu.cn
胡紫娟，北京语言大学信息科学学院，100083，18911419957，18911419957@163.com
王瑞琦，北京语言大学信息科学学院，100083，19801255025，1159925366@qq.com