基于隐性句逗号识别的汉语长句机器翻译

张文娟　　李熳佳　　冯文贺*
广东外语外贸大学外国语言学及应用语言学研究中心/语言工程与计算实验室
{20211050028, 20221050034, wenhefeng@}@gdufs.edu.cn

摘要

长句翻译一直是机器翻译的难题。本文根据汉语中相当数量的逗号（句内标点）和句号（句间标点）可相互转化的特点，提出“隐性句号”（可转化为句号的逗号）和“隐性逗号”（可转化为逗号的句号）概念，并实现其自动识别，以将汉语长句变为短句用于汉英机器翻译。为此，首先通过人工与半监督学习结合方法构建了一个隐性句逗数据集，实现了基于预训练模型的隐性句逗识别方法，其中性能最好的Hierarchical BERT作为后续应用模型。进而，实现了基于隐性句逗识别的汉英机器翻译方法。在WMT2018（新闻）和WMT2023（文学）测试语料上基于预训练机器翻译模型的实验表明，对于汉语长句的英译，本文方法相比基准翻译的BLEU值整体有所提高，而且在相对稳健机器翻译模型上，呈现为句子越长本文方法效果越明显。

关键词： 机器翻译 ； 长句翻译 ； 隐性句逗号

Machine translation of Chinese long sentences based on recognition of implicit period and comma

Wenjuan Zhang　　Manjia Li　　Wenhe Feng
Lab of language engineering and computing
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
{20211050028, 20221050034, wenhefeng@}@gdufs.edu.cn

Abstract

The translation of long sentences has always been a difficult problem for machine translation. In this paper, based on the feature that a considerable number of commas (intra-sentence punctuation) and periods (inter-sentence punctuation) in Chinese text can be transformed into each other, we propose the concepts of "implicit period" (comma that can be transformed into period) and "implicit comma" (period that can be transformed into comma), and realize their automatic recognition to transform Chinese long sentences into short sentences for Chinese-English machine translation. In this paper, a dataset of implicit period and comma is constructed by combining manual and semi-supervised learning methods, and an implicit period and comma recognition method is realized based on a pre-trained model, in which Hierarchical BERT, which has the best performance, is used as the subsequent application model. In this paper,

*通讯作者(冯文贺）：wenhefeng@gdufs.edu.cn

项目基金:广东省社科基金项目（GD24CWY11）；广东省教育厅GK特色创新项目（2023WTSCXO17）；广东外语外贸大学外国语言学及应用语言学研究中心语言与人工智能实验室招标课题（LAI202304）

a Chinese-English machine translation method based on implicit period and comma recognition is realized. The experiments based on pre-trained machine translation models on the WMT2018 (News) and WMT2023 (Literature) corpus show that for the English translation of long Chinese sentences, the method in this paper improves the BLEU value compared to the benchmark translation as a whole, and the effect of the method in this paper is more obvious the longer the sentence is for the relatively robust machine translation model.

Keywords: machine translation , long sentence translation , implicit period and comma

1 引言

篇章翻译是当前制约机器翻译技术性能的一个突出问题，其困难集中体现于长句翻译上 (Koehn and Knowles, 2017)。原因在于长句一般由多个小句（clause)构成，而不同语言间小句及其间结构有重要差异。如表现在汉英语言间，汉语小句无主从之别，长句多流水句，小句间“可断可连” (吕叔湘, 1979)，反映在标点上即逗号（句内标点）、句号（句间标点）可相互转换，句子边界相对模糊；而英语小句主从分明，相互转化性弱，句逗分明，句子边界相对清晰。小句及其间结构与句子边界差异，自然会引发双语长句的翻译问题。长句机器翻译困难也有其计算机制原因。如对于基于循环神经网络的翻译模型，维护长距离依赖关系时可能难以记住大量上下文信息；对于基于注意力机制的翻译模型，会造成注意力分散到更多信息点，导致难以持续聚焦于最关键信息。

对于长句翻译问题，一种解决的思路就是将其化为短句翻译。问题是如何合理将长句化为短句。本文考虑，可从汉语句逗号中有大量可相互转化而不影响原义表达的特点入手，解决汉语长句的翻译问题。例如：

例1 ①少年姓孙，②属马[，]③比小水小着一岁，④个头也没小水高，⑤人却本分实诚。（贾平凹《浮躁》）

E1：①This boy,a member of Sun Family, ②was born in the year of the horse[.]③Although he was a year younger ④and a head shorter than water girl, ⑤he was honest and sincere.（Goldblatt 1991）

E2：①The young Sun, ②a horse, ③is one year younger than Xiaoshui, ④and the individual is no higher than Xiaoshui, ⑤but the person is sincere.（有道翻译2023.11.18）

E3：①The young man’s surname is Sun, ②he was born in the year of the horse, ③and is one year younger than Xiao Shui. ④He is not as tall as Xiao Shui, ⑤but he is honest and earnest.（ChatGPT4 2023.12.10）

（说明：数字序号代表汉英小句序号；汉语隐性句逗号与其对应英译标点用红色[]标出，下同。原文例句、E1及其小句切分采用自(冯文贺, 2019)）

例1汉语复句包含5个小句。语义主题上，小句①②描述少年的个人特征；小句③④和小句⑤虽然也描述少年的个人特征，但却有了对比人物“小水”。据此，前两个小句和后三个小句所表达语义主题是有差异的，其间语义关系也远近有别。也因此，可将小句②后逗号改为句号，而且原文小句间的语义关系远近及具体逻辑语义等不变。本文将此类可转化为句号而不改变原文小句间语义关系远近及逻辑语义的逗号视为”隐性句号”。同理，汉语文本中也存在句号可变为逗号的情况，本文将此类句号视为“隐性逗号”。

将汉语长句变为短句，相比可以实现更好翻译。如专业译者（E1）在小句②后断句翻译，更好翻译表达了源语小句间的语义远近关系，句子的内部结构也更符合英文习惯。相比之下，机器译文E2中句逗号与源语一致，小句间的语义关系不明，句子结构也显得冗长，不符合英文结构习惯。但是，如果不能恰当化长句为短句，也可能导致不良翻译。如机器译文E3中在小句③后断句翻译，根本上改变源语小句间的结构关系，其中割裂了小句③④间密切的并列关系（由连接词“也”表示），隔断了小句⑤对小句③④整体的转折关系（由此也导致译文连接

词“but”不能准确翻译原文连接词“却”的管辖范围）。可以说，长句化短后翻译，在根本上会影响句子内外的结构组织，影响是全局性的。

本文考虑，如果能够识别汉语文本中隐性句逗号，就可能将相当一部分长句经切分重组为较短句子，而经此处理后再进行机器翻译，就可能得到更好的翻译效果。基于此，本文提出了基于隐性句逗号识别的汉语长句机器翻译方法。为此，本文首先构建了汉语隐性句逗号数据集，并实现了其识别模型；进而将隐性句逗号识别模型用于机器翻译的汉语源语数据预处理，以使机器翻译获得更好长句翻译效果。在WMT2018(新闻）和WMT2023（文学）翻译测试集上的实验结果表明，本文方法可以显著提高汉语长句的机器翻译性能，而且长句越长，效果越明显。

本文主要贡献：（1）提出并实现汉语的隐性句逗号识别：针对汉语中句逗号可相互转化现象，提出“隐性句逗号”概念，构建相应数据集，并实现其自动识别技术，为长句化短提供方法。（2）提出并实现基于隐性句逗识别的长句机器翻译方法：结合隐性句逗号识别模型与机器翻译模型，在翻译前将源语汉语长句化短，显著提高了汉语长句的机器翻译质量。

2 相关工作

现有机器翻译模型处理短句表现良好，但对于长句，往往无法给出优质翻译。为改善长句翻译效果，主要进行了两类研究：一类是进行篇章级机器翻译建模，综合解决包括长句翻译在内的篇章翻译问题；另一类是专门针对长句的机器翻译研究。

在篇章机器翻译建模研究中，一般既关注句内词语间结构关系，也关注上下文句子间的一致性、连贯性、结构层次、衔接性等信息(Tan et al., 2019; Chen et al., 2020; Guo et al., 2022; 贾爱鑫et al., 2024)。由于关注更多上下文，一定程度上有利于缓解长句翻译问题。在篇章机器翻译建模中，有研究特别注意到句长的影响。如有研究指出篇章级机器翻译中源语和目标语的句长偏差会导致翻译质量下降；提出动态采样训练数据，以确保不同序列长度可均匀分布；引入长度归一化注意力机制，以使模型聚焦于目标信息，缓解处理较长序列时的注意力偏离问题；提出在解码过程中采用滑动窗口策略，以在不超过最大序列长度的前提下整合更多上下文信息(Zhang et al., 2023)。然而，根本上篇章机器翻译建模并不着重于解决由于（汉英）双语句子边界差异带来的长句翻译问题。

在专门针对长句的机器翻译研究中，一般将长句化为较短的语言单位再翻译。在传统机器翻译（基于规则、基于统计）下，不同语言的翻译上均有研究试图将长句划分较短的语言单位再翻译，一般是利用一些语言特征，如句法模板、结构层次、小句、连接词、标点等，将长句划分为更短的短语、小句、结构片段等后，先翻译再组合(Oliveira et al., 2010; Goh and Sumita, 2011; Yin et al., 2012; Hung et al., 2012; Pouget-Abadie et al., 2014)。在汉英翻译中，也有一些研究尝试利用标点符号、关系代词、层次结构等长句划分为简单句、子句等后再进行翻译(黄河燕and 陈肇雄, 2002; 李幸and 宗成庆, 2006; Xiong et al., 2009)。在神经机器翻译技术框架下，也有研究考虑将长句化短后分别翻译再组合。如有研究在汉英翻译中引入一个拆分和重新排序模型来共同检测源语长句的最佳分割点序列；进而，将每个源语子句由NMT系统独立转换为目标子句，并将翻译的目标子句连接起来形成长句的最终翻译(Kuang and Xiong, 2016)。有研究提出一种提取双语短语的方法来构建短语对齐的双语语料库，并实现了一种长句预处理技术，以此切分长句为短语，以解决长句翻译(Tien and Minh, 2019)。长句分割与句子边界相关，有研究发现发现句子边界分割对口语翻译质量影响显著，提出一种数据增强策略，即在训练过程中将模型暴露于各种边界分割错误中，以提高神经机器翻译系统对句子边界分割错误处理的鲁棒性和机器翻译的准确性(Li et al., 2021)。然而这些研究只是一般地将长句化为较短的语言单位再翻译，而不考虑长句化短后是否改变了源语长句内外，小句间语义关系的远近与逻辑类别等。但事实上，随意切断长句后翻译可能导致原文小句间的逻辑语义结构的改变（如例1的E3）。为此，本文基于汉语部分句逗号可相互转化的特点，提出“隐性句逗号”概念，实现其机器识别，并用以解决汉语长句的机器翻译难题，并特别关注机器翻译中长句内外小句间语义关系的远近与逻辑类别等是否得到了准确翻译。

3 隐性句逗号识别

本文构建了汉语隐性句逗号数据集，并实现了基于预训练语言模型的隐性句逗号识别方法。考虑到原始文本中句逗号的数据不平衡（句号少，逗号多），及隐性句逗号的不平衡，和

类型	汉语文本	英语文本
隐性句号	**S1**祥子，**S2**在与“骆驼”这个外号发生关系以前，**S3**是个较比有自由的洋车夫[，]**S4**这就是说，**S5**他是属于年轻力壮，**S6**而且自己有车的那一类。	he gained the nickname Camel), he was a relatively independent rickshaw man[.]That is to say, he belonged to the young, vigorous set and owned his own rickshaw.
隐性逗号	**S1**从风里雨里的咬牙，**S2**从饭里茶里的自苦，**S3**才赚出那辆车[。]**S4**那辆车是他的一切挣扎与困苦的总结果与报酬，**S5**象身经百战的武士的一颗徽章。	By gritting his teeth through wind and rain and scrimping on food and tea, he finally put enough aside to buy it[,] a tangible reward for his struggles and his suffering, like a medal for valor.

表 1: 人工标注隐性句逗号数据集示例

	数据集划分	数据量	平均长度	总词数
隐性句号	train	15,526	118	1,365,404
	valid	1,940	118	170,785
	test	1,941	115	166,151
隐性逗号	train	11,552	28	438,107
	valid	1,444	27	53,431
	test	1,444	29	55,328

表 2: 半监督扩充数据集统计

	文本长（逗号数）	样本数	字数	平均长度	词表大小
WMT2018 新闻	< 3	34021	1599099	47	3783
	$= 3$	3643	234,953	65	2,834
	$= 4$	1091	82,182	75	2,332
	$= 5$	265	22,961	87	1,692
	> 5	100	10,198	102	1,368
WMT2023文学	> 3	552	34,211	62	1,754

表 3: 机器翻译数据统计

预训练语言模型中本身句逗号知识的不平衡，本文专门构建了一个只包含隐性句逗号的数据集。基于该数据集上，我们训练实现了最优隐性句逗号识别模型，该模型可以相对集中的反映隐性句逗号的特征差异。在机器翻译中，该模型将用于预处理源语汉语文本，由于该模型并未关注真句逗号，其识别结果将与现实文本中句逗号的进行一致性对比调正后作为源语文本预处理结果（见4.1节），输入机器翻译模型进行翻译。

隐性句逗号数据集：为了训练与验证隐性句逗号模型，本文构建了隐性句逗号数据集。首先，在不同体裁的汉语文本上标注共计3000条隐性句逗号；然后通过self-training半监督学习方法大规模扩充数据集。

人工标注由汉语母语者实施，通过两种方式实现。**标注方式1**：标注者根据母语者的语感直接对汉语文本标注。基本判断标准：句逗号相互转变后，语法合理、且不改变原句所含逻辑语义关系的，为隐性句逗号。**标注方式2**：参照汉英翻译标注。标准为：在经典汉英翻译平行语料上，如果英译文本为句号断句，而汉语文本的对应标点处为逗号，则认定该汉语标点为隐性句号；隐性逗号的确定方法同理。具体做法如表 1所示，其中红色标注出了隐性句逗号，连同其左右各一个标点句（用S1、S2等标注）(宋柔, 2022)构成一条数据。如其中的隐性句号数据由S3-S4构成，隐性逗号数据由S3-S4构成。

人工标注共计3000条隐性句逗号样本，其中包含1847条隐性句号标注，1253条隐性逗号标注。随后，采用了self-training半监督学习方法大规模扩充数据集。先将这些标注样本作为初始数据，然后构建了一个基础模型在已有人工标注数据上进行训练，使得模型能够学习到隐性句逗号的标注逻辑和文本特征。接着，利用训练好的模型对未标注数据进行推断，生成伪标签。当模型输出的标签概率高于设定阈值时，将其作为新的标签数据，扩充至初始数据中。通过该方式，最终构建了一个包含33848条数据的隐性句逗号数据集，其中隐性句号19408条，隐性逗号14440条。识别实验中，将该数据集按照8:1:1的比例切分为训练集、验证集和测试集。具体统计结果如表 3所示。

识别模型：本文采用基于预训练语言模型的隐性句逗号分类识别方法。为充分考虑隐性句逗识别中相关语段特征，具体采用Hierarchical BERT模型(Lukasik et al., 2020)对句子对进行编码。如图 1所示，其包含了嵌入层、Sentence Encoder、Context Encoder层，模型输出为隐性句逗号的预测结果。其中Sentence Encoder使用了BERT预训练模型(Devlin et al., 2018)，Context Encoder使用GRU模型。模型的输入为两段文本，分别表示为Sentence1和Sentence2，此处并不是直接拼接两个句子，而是将其层次化并行输入模型。随后Sentence Encoder将学习每个Sentence句内的局部特征并聚合至对应的[CLS]向量中；将两个Sentence的[CLS]表征输入到Context Encoder层中，该层即可赋予文本前后的顺序特征以及

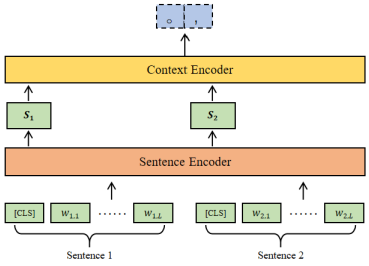

图 1: 隐性句逗号识别模型

模型	accuracy(%)	precision(%)	recall(%)	f1(%)
Hierarchical BERT	97.90	97.82	97.92	97.86
Cross-segment BERT	93.10	92.94	93.05	93.00
BERT-CRF	79.34	96.14	79.34	86.94
BERT-LSTM-CRF	86.17	97.58	86.17	91.52

表 4: 不同模型在隐性句逗号识别上的性能

上下文关系特征，并最终得到文本对的全局特征。全局特征通过线性分类层，将输出映射到隐性句逗的类别标签上，其中线性分类层通过Softmax函数对输出的概率进行了归一化，并利用交叉熵损失函数计算损失。

识别实验：除基于Hierarchical BERT模型外，我们也进行了基于Cross-segmen BERT模型(Wicks and Post, 2021)的方法。与此同时，我们还对比实现了基于序列标注的方法，包括BERT-CRF(Liu et al., 2020) 、BERT-LSTM-CRF(Yang et al., 2022)。实验表明，基于Hierarchical BERT的方法性能最佳。各模型结果见表 4。原因在于Hierarchical BERT相比可以充分学习到句逗号相关语段文本的词汇、语序及上下文特征。而序列标注方法的问题在于，当前的隐性句逗号识别任务下相关数据是独立的文本段，并非实际文本中的句逗号序列。基于Hierarchical BERT的隐性句逗号识别模型将用于后续机器翻译汉语长句化短的预处理阶段。

4 基于隐性句逗号识别的长句机器翻译

4.1 本文方法模型

为验证基于隐性句逗号识别的长句机器翻译方案效果，使用PipeLine方案 (Atrio et al., 2023)进行实现。首先，对源语汉语进行预处理，即进行隐性句逗号识别，并与源语文本比对校正，确定最终句逗号；然后对预处理文本进行机器翻译。如图 2。

隐性句逗号识别：模型所输入的文本为一个句逗号切分的标点句序列（记为sentence1,2,.....n)。为适配本文基于Hierarchical BERT的隐性句逗号识别模型，相邻的两个标点句组合作为一个输入，经过识别模型，预测其间的标点位置为隐性句号或隐性逗号。

句逗校正：由于隐性句逗号识别模型仅考虑了隐性句逗号，而实际文本中为所有句逗号（既包括隐性句逗号，也包括真句逗号），这里需对隐性句逗号模型识别结果进行校正,以获得最终句逗标点。具体做法为，将隐性句逗识别模型输出结果与原文结果进行比对，当模型输出结果与原文一致，保留原文本标点；当模型输出结果与原文本不一致，保留隐性句逗的识别结果。这样做的原因在于，由于受预训练语言BERT自身所包含的大量一般句逗号文本知识的影响，隐性句逗号识别模型并不能完全准确地从真实文本（包含所有句逗号）中识别出隐性句逗号。具体而言，当其标点分类结果与原文本不一致时，可以认为是，句逗模型增强了本文隐性句逗号知识后的结果，即为隐性句逗号；当其与原文一致，可以认为是BERT自身包含的大量真句逗号知识的结果。校正后的句逗标点文本，作为预处理结果输入机器翻译模型。

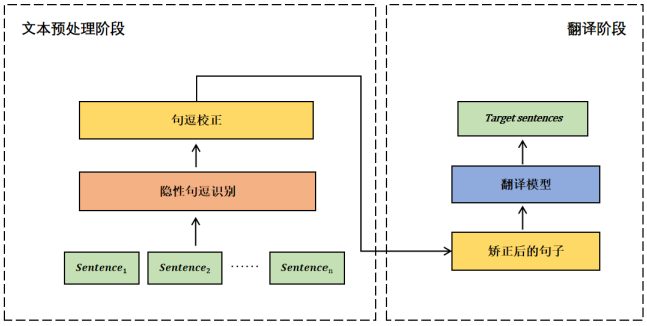

图 2: 基于隐性句逗号识别的机器翻译架构

	数据	文本长（逗号数）	基准翻译	本文方法
Opus-mt	WMT2018 新闻	3 个逗号以下	17.12	16.43
		包含3 个逗号	15.88	16.74
		包含4 个逗号	15.14	16.11
		包含5 个逗号	15.30	16.58
	WMT2023 文学	3 个逗号以上	16.42	17.11
Randeng-mt	WMT2018 新闻	3 个逗号以下	18.42	17.93
		包含3 个逗号	15.88	16.75
		包含4 个逗号	14.65	15.24
		包含5 个逗号	13.93	14.39
	WMT2023 文学	3 个逗号以上	14.29	14.18

表 5: 翻译结果比较

4.2 实验设置

数据集：为验证本文方案效果，本文在标准机器翻译任务WMT2018的中英平行翻译数据集（新闻）[1]和WMT2023中的中英文学翻译平行语料[2]的测试集上进行基于隐性句逗号识别的长句机器翻译实验。引入文学语料是考虑，文学文本是机器翻译难点，而且文学文本相比一般文本口语性强，可逗可句或可断可连的情况更突出，特别适于本文方法的验证。为了充分验证本文方法在汉语长句上的效果，进一步按照文本长度（以包含的逗号数计算）将数据集划分如表 3。考虑是，逗号是句内标点，一个文本包含的逗号数越多，往往意味着其中的句子包含的小句越多，句子越长。

翻译模型：在翻译阶段，本文使用两个表现优异的预训练机器翻译模型进行了实验，分别是opus-mt (Tiedemann and Thottingal, 2020)和Randeng-mt (Zhang et al., 2022)。二者均基于transformer的端到端架构，并基于大型中英平行语料库训练，包含了丰富的双语知识。实验中二者具体设置相同，主要是：设置束搜索宽度为2；采用多概率采样；top_k为40，其可以在生成过程中过滤掉不太可能的词，仅保留头部的40个词；top_p概率为0.8，其代表输出词的概率之和达到或超过0.8，才会在生成过程中保留；此外为确保模型不重复地连续输出，限制了模型连续重复生成单词的个数为5。

评估指标：采用通用的BLEU值 (Papineni et al., 2002)。其主要通过比较机器翻译的输出和参考译文的n-gram 相似度计算得出。BLEU值中的n-gram的取值为4。

[1]https://www.statmt.org/wmt2018/translation-task.html
[2]https://aclanthology.org/events/wmt-2023/2023wmt-1

样例	类别	文本内容
1	原文	**S1**四是织密织牢民生保障网，**S2**增进人民福祉。**S3**我们坚持以人为本，**S4**持续增加民生投入[，] **S5**保基本、兜底线、建机制，**S6**财政用于民生的比例达到70%以上
	原文翻译	Four are the well-being of the people through the woven network of secure livelihoods. In order to promote the well-being, we insist on people-centred and sustained increases in human investment[，] and the preservation of basic, bottom-line and institutional mechanisms, the share of the financial resources spent on human life is over 70 per cent.
	修正文本	**S1**四是织密织牢民生保障网，**S2**增进人民福祉。**S3**我们坚持以人为本，**S4**持续增加民生投入[。] **S5**保基本、兜底线、建机制，**S6**财政用于民生的比例达到70%以上
	修正文本翻译	The fourth is to weave a network of secure livelihoods for the betterment of the people. To improve the well-being of the people, we insist on people-centred efforts to increase people's inputs[.] To protect the basics, to get the bottom line, to build the mechanisms, the proportion of the funds spent on people's lives is more than 70%.
2	原文	**S1**而此时还处在大厅里的人，**S2**都并不急于领取任务[，] **S3**他们都认为顾飞一定会来自首[。] **S4**而自首的地方同样是这个大厅，**S5**因此这些人都紧盯着进进出出的玩家[，] **S6**一发现是法师就两眼放光，**S7**接着又是一脸失望。
	原文翻译	Many of the people inside the hall did not immediately receive the mission[ ] as they were confident that Gu Fei would turn himself in [.] They closely observed all the players coming and going since this hall was also a place where fugitives could turn themselves in [.] Their eyes would periodically gleam whenever they saw a Mage and then be followed by a look of disappointment.
	修正文本	**S1**而此时还处在大厅里的人，**S2**都并不急于领取任务[。] **S3**他们都认为顾飞一定会来自首[，] **S4**而自首的地方同样是这个大厅，**S5**因此这些人都紧盯着进进出出的玩家[。] **S6**一发现是法师就两眼放光，**S7**接着又是一脸失望。
	修正文本翻译	And no one in the hall is in a hurry to get a job [.] They all think that Luo Fei will turn himself in[,] and that the same hall is where they turn themselves in, so they're all staring at the players who come in and out [.] The wizards are looking at the light, and then they're disappointed.

表 6: 原文及译文与修正文本及译文对比

4.3 实验结果

如表 5所示：

（1）随着逗号增多，即句子包含的小句数增多，句长增大，机器翻译效果变差，充分证明长句越长对机器翻译的挑战越大。

（2）在包含3到5个逗号的语段文本内，本文方法比基准方案的翻译质量整体有所提高，其中在更稳健翻译系统（Opus-mt）上，本文方法的BLEU值分别提高了0.86、0.97和1.28，也即随着源语文本长度的增加，本文方案的优势变得更明显。

（3）本文方法也适应于文学翻译，在Opus-mt系统上，本文方法比基准模型提升了0.69个BLEU值。一般认为文学翻译难度大，主要是文学文本中更多人物对话和叙事，也更多涉及日常生活，内容容易理解，但也因此句子口语性强，结构更灵活，可断可连的句子更多，句子边界相比更模糊，翻译断句难度更大。本文方法可为文学机器翻译难题的解决提供一种特别思路。

（4）值得注意，本文方法对于相对较短的句子（包含3个以下逗号的语段文本）翻译，并没有体现出优势。这一方面反映出短句翻译并非机器翻译难题，另一方面也反映出本文方法也还未能更好识别真句逗号。这是因为隐性句逗号识别模型仅专注于隐性句逗号的识别，而暂时忽视了真句逗号的问题。这无疑需要在进一步工作中予以合理解决。

4.4 实例分析

表 5分别给出了汉语原文与其机器翻译结果和经本文方法修正标点后的汉语修正文本与其机器翻译结果。

实例1 文本选自WMT2018的新闻文本。修正文本将S4后的逗号改为了句号，修正后S3、S4、S5、S6间的关系远近更清晰，逻辑语义更明确。对比修正文本翻译与原文翻译可以发现，修正文本翻译更好地反映了S3与S4的关系，S5与S6的关系，而且相比原文翻译语法

结构也更合理。

实例2 文本选自WMT2023的文学文本。修正文本将S2后的逗号改为了句号，将S3后的句号改为了逗号，将S5后的逗号改为了句号。修正后S3、S4、S5间的紧密关系得以凸显，包括S3与S4的递进性关系（都是“他们都认为”的内容，并且用“而”连接），S5与S3、S4间的因果性关系（由“因此”体现）。并S3-5与S1-2和S6-6的关系距离也相对拉开，显得更清晰。对比修正文本翻译与原文翻译，修正文本翻译较好地反映了各个S间关系的语义远近与逻辑关系关系。

5 总结

针对长句机器翻译难题，本文提出基于隐性句号逗识别的汉语长句机器翻译方法。本文首先构建了隐性句逗号号数据集，并基于预训练模型实现了隐性句逗自动识别；进而将隐性句逗号识别模型作为翻译数据预处理模块与翻译模型结合，以解决汉语长句翻译难题。实验结果显示本文方法对长句的翻译性能有显著提升，而且长句越长效果越佳；这一方法也提升了文学翻译的长句翻译性能。未来我们将进一步完善隐性句逗号识别方法，特别是完全考虑真实语境中所有句逗号（真实句逗号与隐性句逗号）的识别问题，并将探索直接在机器翻译模型中融入隐性句逗号的理解。

参考文献

Àlex R Atrio, Alexis Allemann, Ljiljana Dolamic, and Andrei Popescu-Belis. 2023. A simplified training pipeline for low-resource and unsupervised machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 47–58.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. *arXiv preprint arXiv:2006.04721*.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*.

Chooi-Ling Goh and Eiichiro Sumita. 2011. Splitting long input sentences for phrase-based statistical machine translation. In *The Association for Natural Language Processing*, pages 802–805.

Jianming Guo, Xinran Chen, Zihan Liu, Weijie Yuan, Jianshen Zhang, and Gongshen Liu. 2022. Context modeling with hierarchical shallow attention structure for document-level nmt. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. 2012. Sentence splitting for vietnamese-english machine translation. In *Proceedings of the 2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 156–160.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation*, pages 28–39.

Shaohui Kuang and Deyi Xiong. 2016. Automatic long sentence segmentation for neural machine translation. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 162–174. Springer.

Daniel Li, I Te, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield. 2021. Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7553–7557. IEEE.

Mingyi Liu, Zhiying Tu, Zhongjie Wang, and Xiaofei Xu. 2020. Ltp: a new active learning strategy for bert-crf based named entity recognition. *arXiv preprint arXiv:2001.02524*.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Francisco Oliveira, Fai Wong, and Iok-Sai Hong. 2010. Systematic processing of long sentences in rule based portuguese-chinese machine translation. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010*, pages 417–426.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, Kyunghyun Cho, and Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 78–85.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. 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 1576–1585.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Ha Nguyen Tien and Huyen Nguyen Thi Minh. 2019. Long sentence preprocessing in neural machine translation. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007.

Hao Xiong, Wenwen Xu, Haitao Mi, Yang Liu, and Qun Liu. 2009. Sub-sentence division for tree-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 137–140.

Chongluo Yang, Long Sheng, Zhongcheng Wei, and Wei Wang. 2022. Chinese named entity recognition of epidemiological investigation of information on covid-19 based on bert. *Ieee Access*, 10:104156–104168.

Bao sheng Yin, Junjun Zuo, and Na Ye. 2012. Long sentence partitioning using top-down analysis for machine translation. In *Proceedings of IEEE CCIS2012*, pages 1425–1429.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv:2209.02970*.

Zhuocheng Zhang, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556.

冯文贺. 2019. 汉英篇章结构平行语料库构建与应用研究. 科学出版社.

吕叔湘. 1979. 汉语语法分析问题. 商务印书馆.

宋柔. 2022. 小句复合体的语法结构. 商务印书馆.

李幸and 宗成庆. 2006. 引入标点处理的层次化汉语长句句法分析方法. 中文信息学报, 20(4):8–15.

贾爱鑫, 李军辉, 贡正仙, and 张民. 2024. 融合目标端上下文的篇章神经机器翻译. 中文信息学报, 38(4):59–68.

黄河燕and 陈肇雄. 2002. 基于多策略分析的复杂长句翻译处理算法. 中文信息学报, 16(3):1–7.