基于大语言模型的汉语主宾可逆句语义与施事程度评估

刘柱
清华大学人文学院中文系
liuzhu22@mails.tsinghua.edu.cn

摘要

主宾可逆句是现代汉语语法体系中较为特殊的语法现象，其基本特征是句子中的主语和宾语可以相互交换位置而不影响基本句义。换句话说，交换成分的语义角色并未发生明显改变，从而产生了“格配置变动”。与之相对的主宾不可互逆句则往往由于论元与核心动词的典型施受关系，交换主宾语后语义发生反向改变。大语言模型在大规模语料上进行训练，并取得了卓越的文本理解能力。一个值得探究的问题是，它能否正确区分这两种情况，以及理解背后的施受语义关系？本文收集了相关的语料，并对现有的汉语大语言模型进行句子等义性和各个成分的施事度进行评估。并得出如下结论：1）一般取末层所有词向量的平均方式不足以区别两种情况；2）通过采用不同层的信息，大语言模型可以在一定程度上反映不同位置上的施事程度。

关键词： 主宾可逆句；施事程度；大语言模型；评估

Evaluation of the semantics and agential degree of Chinese subject-object reversible sentences based on large language models

Zhu Liu
School of Humanities, Tsinghua University
liuzhu22@mails.tsinghua.edu.cn

Abstract

The subject-object reversible sentence is a relatively special grammatical phenomenon in the modern Chinese grammar system. Its basic feature is that the subject and object in the sentence can exchange positions without affecting the basic sentence meaning. In other words, the semantic role of the exchange components has not changed significantly, resulting in a "case configuration change". In contrast, sentences with irreversible subject and object often have reverse semantic changes after exchanging the subject and object due to the typical giver-recipient relationship between the argument and the core verb. Large language models are trained on large-scale corpora and achieve excellent text understanding capabilities. A question worth exploring is whether it can correctly distinguish between these two situations and understand the semantic relationship between give and take behind it? This article collects relevant corpus and evaluates the existing Chinese large language model for sentence equivalence and the implementation degree of each component. And the following conclusions are drawn: 1) Generally, the average method of all word vectors in the last layer is not enough to distinguish between the two situations; 2) By using information from different layers, large language models can reflect the degree of action at different positions to a certain extent.

Keywords: Subject-object reversible sentences , agential degree , large language model , evaluation

1 引言

主宾可逆句，又称主宾互易句、主宾易位句、双面句式等，是指在不增减任何词语的情况下，动词左右的主宾语发生交换而语义保持基本不变的现象(张玲娟, 2012)。例如，“十个人吃一顿饭 $\Longrightarrow$ 一顿饭吃十个人”。这时，变易后的相同成分的语义角色往往不发生改变（“十个人”在前后变化前都是施事者，“一顿饭在前后都是受事者。），也由此母语者不会产生语义混淆。与之相对的是主宾不可逆句，即主宾交换位置之后，语义完全相反。例如，“大鱼吃小鱼 $\not\Longrightarrow$ 小鱼吃大鱼”。由于汉语的基本语序为“主语-动词-宾语”（SVO, Subject-Verb-Object）并且与之对应的典型格配置为“施事-动作-受事”，典型场景下应该是主宾不可逆句。由于主宾可逆对和主宾不可逆对变换的形式完全一样，语义变化却产生了相反的效果，因此对于汉语的二语学习者或者计算模型而言，区分这两种情况都富有挑战。

理解句对的同义性对于目前的计算语言模型非常重要。最新的大语言模型，例如ChatGPT (OpenAI, 2023), CPM-Bee [0]在海量文本上进行了训练，并在句子级别的多个任务上都取得了卓越的表现，例如自然语言推理任务、句对匹配任务(Zhao et al., 2023)等。其中这些任务都涉及到判断两个句子是否同义，这就自然引发一个研究问题，即“大语言模型是否可以判断出主宾可逆和不可逆两种情形下的句对语义变化？”这一任务对计算模型存在挑战：（1）由于不可逆句是汉语典型的格配置场景，涉及该类型的句子自然分布更广，而模型倾向于学习到与训练句子中相似的分布，因此可逆句的判断更加困难；（2）施事性等语义角色是数据（即句子）的隐性知识，模型从表面的形式学习到这种深层理解则需要模型强大的泛化能力。

受此启发，我们针对这一问题进行了实验设计。首先收集不同场景下的语料，即可逆句对和不可逆句对。其次，本文对现有的中文通用大语言模型和专用语言模型在不同类型下的句对语义相似性做了评估，并通过相关系数观察模型是否可以反映正确的语义相似性。其次，设计了反映各成分施事程度的语料库级别的指标，并计算了不同层数的输出对于结果的影响。最后，本文通过不同案例进一步分析了大语言模型对施事程度的反映，以此挖掘模型潜在的判别能力。实验表明，1）以往利用句子内所有词向量的平均作为句向量的方式不足以区别上述两种情况；2）不同层对于施事程度的判断不尽相同，这反映出不同阶段的输出对于词序的敏感度不同；3）整体上看，大语言模型可以在一定程度反映不同成分的施受事程度。

2 相关工作

2.1 主宾可逆句与语义角色

主宾可逆句是汉语中的一种特殊句式，并且被不同的研究者所重视。宋玉柱 (宋玉柱, 1991)、任鹰 (任鹰, 1999)、李宇明 (李宇明, 2002)、等都提出过这一特殊的句式。对于可逆句的类型描写，则以宋玉柱 (李宇明, 2002)和李敏 (李敏, 1998)的工作为代表。其中宋玉柱根据句型表达的语法意义差别将其分为三类：供动型、被动型和从动型。李敏则从动词语义入手提出了六种分类类型：混合义、依附义、供给义、笼罩义、充满义和进入义。同时主宾互逆收到很多限制因素，包含句法和语义两方面 (李宇明, 2002)。

2.2 计算模型的评估和知识对齐

大规模语言模型是自然语言处理最新的研究成果，它通过扩展自身的模型规模、并在海量数据上进行训练，从而达到理解和生成人类语言的通用能力(Zhao et al., 2023)。比较典型的模型有仅基于解码器模型的GPT系列 (OpenAI, 2023)，仅基于编码器的BERT (Kenton and Toutanova, 2019)，以及序列到序列生成的T5模型 (Raffel et al., 2020)。大语言模型在各种评测数据集上取得了很好的准确性。很多研究者开始评估其他各个方面是否与人类的知识对齐 (Ji et al., 2023)，例如语言能力，幻觉问题，逻辑推理能力,安全性和隐私保护等。全面、可靠的评估是可解释人工智能的必由之路。

[0]https://github.com/OpenBMB/CPM-Bee

3 研究方法

我们首先确定研究对象的范围，并据此收集相关语料，之后我们对模型进行评估，包括模型对于成对语句是否同义的判断以及施受事的反映程度。

3.1 研究对象

假设$\mathcal{S}$代表由所有语法表达正确、语义明确的简单小句组成的空间，并且该集合的句子成员$s \in \mathcal{S}$，仅由主语A，宾语O，以及连接它们的动词V构成。这些在句子s中的成员可以表示为s^A, s^O和s^V，该句子简记为$s = AVO$，并且对应于该形式句子s的语义可以表示为：$m(s)$。我们同时定义一个主宾部分可逆的操作，即$s' = OVA$。在$\mathcal{S}$空间中，我们关注两类子空间，一类称之为主宾可逆句集合$I \in \mathcal{S}$，另一类称之为主宾不可逆集合$U \in \mathcal{S}$。这两类集合的定义如下：

$$\begin{aligned} I &= \{s \in \mathcal{S} | m(s) = m(s')\} \\ U &= \{s \in \mathcal{S} | m(s) \neq m(s')\} \end{aligned} \tag{1}$$

值得注意的是，可逆操作对于I和U两个空间(即，由它们构成的整个空间$\mathcal{S}$)都具有对称性，即这两个空间中的任何一个元素在进行了可逆操作之后，它仍然处于原空间。用公式表达为：

$$\{s' | s \in \mathcal{S}\} = \mathcal{S} \tag{2}$$

换句话说，互为可逆的句子之间没有方向性。出于为了研究的方便，本文规定了其中的方向性，其中的源句的S具有较强的施事性，其语义角色可以由主格充当；O具有较强的受事性，其语义角色可以由宾格或者旁格构成，而目标句则为源句可逆操作后的形式，我们用箭头表达其中的方向性，即

$$\begin{cases} s \implies s' & \text{if } s \in I \\ s \not\Rightarrow s' & \text{if } s \notin U \end{cases} \tag{3}$$

3.2 评估方法

本文首先评估模型f是否能反映可逆句和不可逆句的语义变化，过程如下：首先模型以由N个词汇w_i构成的句子$s = \{w_1, w_2, ..., w_N\}$作为输入可以针对每个词汇输出一个$d$维的隐向量$\{h_1, h_2, ..., h_3\} = f(\{w_1, w_2, ..., w_N\} \in \mathbb{R}^d$，之后可以通过不同的方式可以获取代表句子语义的句向量，常见的有选择均值$\mathbf{e} = \frac{1}{N}\sum_{i=1}^{N} h_i$(Xu, 2023)和选取最后一个词汇$\mathbf{e} = h_N$的方式，后一种方式尤其适合于基于自回归预测任务的生成式大语言模型(OpenAI, 2023)。之后，比较可逆句对和不可逆句对之间的相似性，来确定模型是否可以正确反映它们的语义变化，我们期待可逆句对的平均相似性要显著大于不可逆句的相似性，即

$$\frac{1}{N}\sum_{s \in I} Sim(\mathbf{e}_s, \mathbf{e}'_s) > \frac{1}{M}\sum_{s \in U} Sim(\mathbf{e}_s, \mathbf{e}'_s) \tag{4}$$

同时，为了和之前的结果有一定的可比性，我们将可逆句对的真值相似性标签设置为1，不可逆句的真值相似性标签设置为0，通过输出真值标签和模型输出结果相似度$Sim(\mathbf{e}_s, \mathbf{e}'_s)$的斯皮尔曼相关性系数，来进一步反映模型的表现性能。由于可否可逆与主宾语的施受性相关，我们定义了某一成分的施事度P这一指标，施事度越高，该发出者施加的作用越强，往往对另一方受事者产生直接的、重大的影响，与之相应地，该受事者的施事度也达到最低，而该施事的动作也施动性更强。而相反，施事度越低，其施加给另一方的影响越弱，越弱到被影响则变为了受事。一般的格配置会讲主格编码为施事度最强的成分，旁格则次之、宾格则施事度最弱，一般为受事。由于语义的需要，不同施事度的成分也会出现升降格的现象。我们认为来自不同可逆配置的不同成分具有如下的施事度由强到弱的排序，即

$$A/V_U > A/V_I > O_I > O_U \tag{5}$$

相应地，我们通过句对之间相应位置的成分相似性在可逆和不可逆、不可逆两种配置中施事度的差异来反映模型对于这一特性的把握。该成分相似性使用相应位置单词的输出向量的均值来表示，相应的类型和举例参考表 1。其中“左左”表示，（不）可逆的源句左成分和可逆操作后的目标句的左成分的相似性，其他类型可类推。两种配置下的相似性差异见最后一列。

类型	不可逆成分	可逆成分	不可逆>可逆
左左	(大鱼，**小鱼**)	(一顿饭，**十个人**)	是
右右	(小鱼，**大鱼**)	(十个人，**一顿饭**)	是
左右	(大鱼，**大鱼**)	(十个人，**十个人**)	否
右左	(小鱼，**小鱼**)	(一顿饭，**一顿饭**)	否
中	(吃，**吃**)	(吃，**吃**)	是

Table 1: 不同位置类型在是否可逆的两种配置下的相似性程度差异以及举例说明

模型	可逆句	不可逆句	句内随机	句间随机	相关性	其他数据结果
SBERT	93.2	92.0	88.5	71.2	43.0	-
CoSENT	96.0	93.8	91.7	56.2	**51.1**	63.1
CPM(M)	79.4	91.3	58.9	26.6	41.1	-
CPM(L)	46.4	60.0	41.3	25.0	26.3	-

Table 2: 不同模型在不同类型的数据下句子对之间的平均相似度

4 实验过程

4.1 收集语料

本次实验采用的可逆句数据来源于论文 (李敏, 1998)，该论文根据动词的语义对可逆句分为了不同的类型，包括“混合义”、“依附义”、“供给义”等六种类型，最终共收集了44条可逆句。对于不可逆句，本文首先选择词频较高的动词，内省创造出20条不可逆句。这些句子可以参考附录 7。

此外，我们使用随机打乱了句内字和无关句子作为对照，分为命名为“句内随机”和“句间随机”，其数量与可逆句数量相当。

4.2 语言模型

本文采用的大语言模型为CPM模型，它使用万亿级高质量中英语料进行预训练，并在中英双语的常见任务中取得优异成绩。其架构采用48层Transformer，本文使用平均方式和最后一个词向量的方式来表示句向量，分别使用CPM(M)和CPM(B)来表示。作为比较模型，实验还采用了开源句子工具包 (Xu, 2023)中的模型，包括CoSNET模型[1]和句向量表示模型SBERT[2]。前者采用有监督的方式训练BERT和Softmax分类函数，后者使用一种基于排序的损失函数。其中CoSNET模型在公开评测的数据集上都取得更优的性能。值得注意的是，它们在评估句向量时，均采用最后一层词向量求均值的方式，我们将它作为一种默认的方式。

本实验仅使用最终训练好的模型进行句向量评估，不涉及模型的训练，其他推断的参数都采用默认的方式。

5 结果分析

我们主要分析了三种情况，第一种是不同模型在不同配置下的相似性。第二种是大语言模型配置下，不同层数的输出对于五种位置的施事度的影响。最后一类情况是在最优的层数配置下，不同类型样本的施事度分布情况以及案例分析。

5.1 相似度分析

表 2罗列出来了四个模型在不同类型数据下变换句子前后的相似度，并在整个样本维度上取了平均。可以看出通过常规方式（最后一层、词向量取平均）得到的句向量，没办法区分可逆和不可逆的情况，尽管SBERT和CoSENT在可逆句上取得很大的相似度，但其在不可逆句和句内随机上却取得了相比的高相似度，这说明它对汉语字词的位置信息很不敏感，打乱了词序

[1] https://huggingface.co/shibing624/text2vec-base-chinese-paraphrase
[2] https://huggingface.co/DMetaSoul/sbert-chinese-general-v2

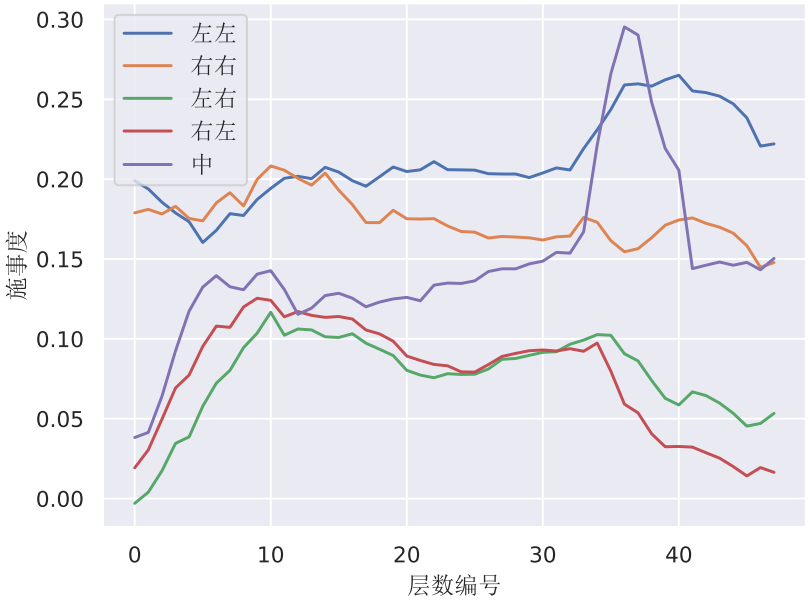


Figure 1: 层数对于各个位置类型的施事程度的影响

位置类型(%)	可逆句	不可逆句	施事度	p值
左左	38.0	64.5	26.5	5.54e-6
右右	33.5	54.3	20.8	1.86e-6
左右	74.6	63.0	11.6	1.03e-5
右左	72.5	59.9	12.6	1.84e-5
中	56.7	86.2	29.5	3e-4

Table 3: 五种位置类型下两种句型的施事度比较

的句子语义对它理解并不造成很大的困扰；而对于CPM大语言模型，不仅仅可逆句没有取得很好的相似性，不可逆句的相似性比可逆句的甚至还要高出很多，相应的它们与真值标签的相关性也很低。不过它对于随机排列的句子的相似性判断更好，暗示它可能对于位置信息更加敏感。相关性最高的CoSENT模型也远低于它在其他公开数据集的平均表现，这说明模型对于汉语可逆的语义相似判断仍旧是一个有困难的任务，同时常规的句向量获取模型可能存在偏差。

5.2 层数影响

我们接着分析在不同层数后选取的向量在不同位置类型下模型对于施事度的把握，施事度的定义可以参考 3.2。图 1展示了这一结果。可以发现，层数对于施事度的影响趋势在不同的位置下不同，但所有配置下的平均施事度都大于0，这反映了模型还是可以一定程度地区分不同结构中的施受性变化的。而对于每种情形下最佳层数的施事度而言，动词上反映的也越强烈（0.450），这暗示可以通过施事度对某些动词进行进一步的细分。

5.3 施事度分析

之后，我们输出最佳层数配置下的样本的施事度分布，结果见图 2。可以看出各种情况下的施事度关系都符合预期。表 3也输出了可逆句和不可逆句在相应情况中的相似度均值、施事度和差异p值。施事度都为正数，且p值很小，这表明模型可以分别出不同位置下的施事程度。

我们分析了在动词比较下不可逆句相似的前五个例句（仅展示源句）和可逆句最不相似的五个例句，见表 4。可以看出，不可逆句最相似的一些动词出现在情感动词上，并且它的主宾语相互之间是平等的，这样使得动词在互换的两句话之间几乎没有太大的变化。而可逆句中最

不可逆句	可逆句
我想念我的朋友们	桌子蒙着一块花布
总理爱人民	太阳晒着稻草堆
下属害怕领导	鲜花开遍了原野
小蝌蚪找妈妈	空格里签名字
学生听老师	茄子炒肉

Table 4: 动词相似性前五（不可逆）和后五（可逆）的源句

不相似的动词很多具有“覆盖”义，这类动词不能很清晰地判断施受关系，仅通过位置可以大略地体会出来，因此交换位置后的句子，动词的相似性也会随之减弱。

6 结论

本文探讨了大语言模型对于汉语中的可逆句和不可逆句的语义判断情况，并进一步分析了模型对于不同位置的施事度是否在两类句子中存在差异。实验结果表明通过一般方式得到的语言模型的输出不能很好区分这两种情况。同时，大语言模型可以比较好地反映不同位置的施事性特征，这提醒我们模型可能具理解这类及物小句的潜能，未来可以收集更加全面、自然的语言数据，继续探究如何更好地表现语义。

参考文献

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ming Xu. 2023. Text2vec: Text to vector toolkit. https://github.com/shibing624/text2vec.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

任鹰. 1999. 主宾可换位供用句的语义条件分析. 汉语学习.

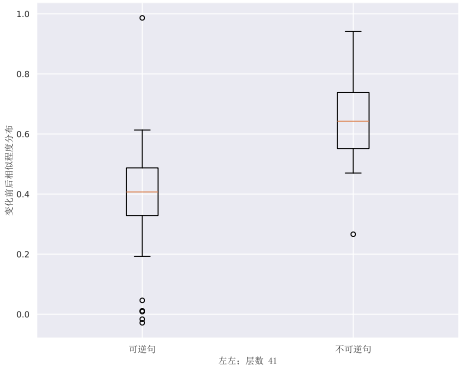
宋玉柱. 1991. 可逆句现代汉语特殊句式. 江西教育出版社, 南昌.

张玲娟. 2012. 现代汉语主宾互易句研究. Master’s thesis, 山东大学.

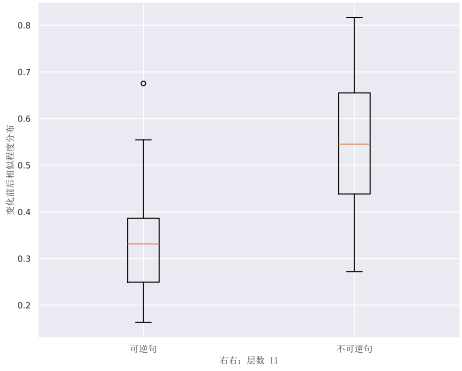
李宇明. 2002. 存现结构中的主宾互易现象研究. 商务印书馆.

李敏. 1998. 现代汉语主宾可互易句的考察. 语言教学与研究.

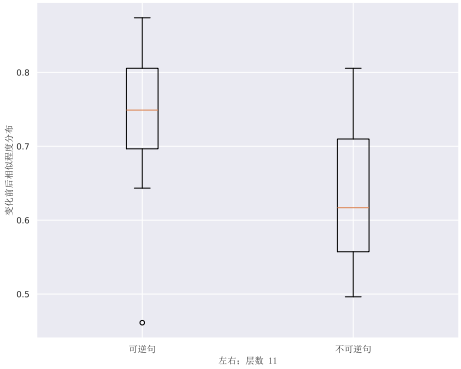
7 附录



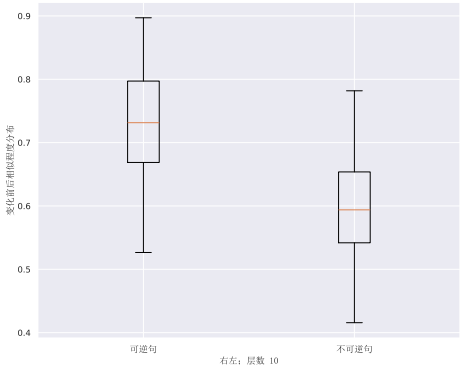
(a) 左左



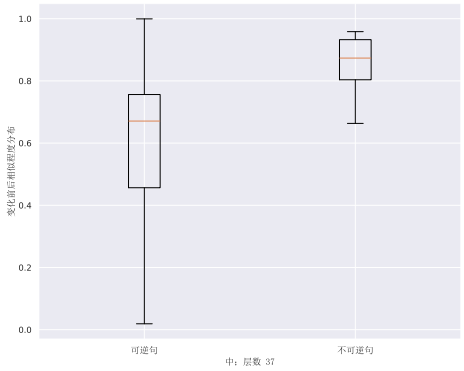
(b) 右右



(c) 左右



(d) 右左



(e) 中间

Figure 2: 不同位置类型下的可逆句和不可逆句变化相似性分布

可逆句	不可逆句
两份水泥配一份沙子	一匹马抬两个人
鱼头炖豆腐	两个人搬一张沙发
白菜熬粉条	两个人抬一张桌子
鸡蛋炒黄瓜	总理爱人民
小葱拌豆腐	张三追累了李四
茄子炒肉	秦国打败了燕国
好苹果放上面	老师打了学生
菜摆桌子上	那个女生笑话他
粮食堆仓库里	学生听老师
像片贴右上角	赵国游说秦国
名字签空格里	大鱼吃小鱼
口袋缝左边	我想念我的朋友们
玉米种前院	下属害怕领导
地瓜种后院	他刚失去了她
三个人住一个屋子	后排的人猛推前面的人
两个人骑一匹马	他拉了后面的人
五个人坐一条板凳	小蝌蚪找妈妈
两个人睡一张床	同桌踢了他
七个人吃一顿饭	哥哥保护弟弟
好几个人洗一盆水	家长批评孩子
大楼笼罩着晨雾	
大地覆盖着白雪	
山谷弥漫着烟雾	
稻草堆晒着太阳	
汽车盖着油布	
桌子蒙着一块花布	
天空布满了乌云	
原野开遍了鲜花	
前沿布满了地雷	
大地洒满雪花	
天空飞满了树叶	
街头聚满了人群	
商场挤满了顾客	
屋里已经进了不少水	
我们班又插进了三个女生	
游行队伍里夹进了一个便衣警察	
暗房里透进一线光亮儿	
他的血管里输进了二百毫升的人造血浆	
人造血浆输进了他的血管里	

Table 5: 可逆句和不可逆句示例