文章编号：1003-0077（2011）00-0000-00

关于句末标点符号对口语机器翻译调序的影响的研究*

王明轩，吴晓湘，熊浩，刘群

（中国科学院计算技术研究所，北京市 100190）

摘要：与书面语相比，口语有更复杂的语气成分，比如疑问句或者祈使句等。其中，疑问句在口语表达中占据了重要的作用。但是目前的翻译模型在处理疑问句翻译上有明显的不足，因为目前还很难解决长距离调序问题，而表达语气的信息通常在句尾 1，但是需要调序的部分通常在句首 2，所以目前的机器翻译模型很难在调序的时候充分利用到句子的语气信息。针对这种情况，本文提出了一种在训练语料中将句尾标点符号提前的方法，通过将句尾语气信息复制到句首，使机器翻译模型能准确的捕捉到句子的语气信息，从而提高翻译的准确度。在大规模汉英口语翻译实验中，使用该方法对互联网口语语料进行处理后，翻译效果有了 2.49 个 BLEU 值的显著提升。

关键词：口语翻译；标点信息；机器翻译

中图分类号：TP391　　**文献标识码：**A

The Research of the Influence of Punctuations in the End of Sentence on the Oral Machine Translation when Doing Reordering

Wang Mingxuan, Wu Xiaoxiang,Xiong Hao ,Liu Qun

(Institute of computing technology, Beijing China, 100190)

Abstract: Compared with written language, oral language has more complex tone ingredients, such as interrogative or imperative sentences and so on. Among them, interrogative sentences play an important role in oral language. However, there are obvious deficiency in the current English-Chinese translation model when dealing with interrogative sentences because it is still very difficult to solve the problem of long-distance reordering. It is quite usual that tone information is located in the end of the sentence, but the part needed to be reordered is usually at the beginning of the sentence, so it is hard for the current translation model to fully utilize the tone information in the sentence when doing reordering. For this situation, we propose a method to put the punctuations to the beginning of the sentence, which can bring the tone information expressed in the end of sentence to the beginning and thus improving the quality of translation. We apply this method to large scale experiments on Chinese-English oral translations, and eventually we significantly improve the performance in terms of BLEU by 2.49.

Key words: Oral Translation tone Information; Statistical Machine Translation

1 引言

疑问句的翻译效果对口语翻译的质量起着至关重要的作用，但是目前大多数翻译模型对疑问句的翻译存在很多问题。这是因为英语疑问句的结构与汉语相比有很大的不同。比如“明天是星期天吗?” 和“Is tomorrow Sunday?”，汉语的疑问句一般只是在陈述句中末尾简单的加入疑问词，而英语疑问句与汉语相比需要将疑问词提前到句首或者在句首新增加疑问词。目前大部分解码模型对于这种需要长距离调序的句子，处理的不是很好，以汉英口语翻译为例，对于疑问句的翻译通常会出现下面两种错误：

调序错误　疑问句的翻译只是简单的在陈述句末尾添加问号，比如“这周你准备做什

* 本文受 863 重大项目课题 2011AA01A207 对本文研究工作的资助。

* **作者简介：**王明轩：男，1989 年生；硕士研究生，研究方向是自然语言处理、机器学习。
吴晓湘：男，1991 年生；本科，研究方向是自然语言处理。
熊皓：男，1985 年生，助理研究员，研究方向是自然语言处理、机器学习。
刘群：男，1966 年生，研究员，研究方向是自然语言处理、机器翻译、信息提取。

么?” 翻译成“This weekend you ready to do what?”。

疑问词丢失 有一些仅仅表示语气信息的虚词在翻译的时候丢失，这种现象比较多，比如“你住在哪里？” 被翻译成“Where you live?”，缺少了虚词，正确的翻译应该是“Where do you live?”，“do” 在这里没有具体的含义，很多时候很难在目标端被正确的翻译出来。

针对上面提到的这两种情况，我们考虑可以将语料中句尾的标点符号都复制到句首，这样翻译模型就可以比较准确的捕捉到句子的语气信息，对于英语疑问句在句首的调序更加准确从而对疑问句的翻译有很大的提升。比如训练语料中有这样的句子：

你帮助他们打扫房子吗?

Do you help them to clean the room ?

用这样的句对来训练来模型，翻译类似的句子“你有什么专门要参观的吗?”，句尾的信息很容易在翻译前面的句子的时候被忽略，利用层次短语模型这个句子被翻译成了“there anyting special you want to visit ?”。因为距离太远了，利用语言模型约束对距离近的词语效果比较好。如果把训练语料和开发测试集句尾的问好复制到句首，抽规则或者用语言模型约束的时候都会用到这个信息。如果采用我们的方法对训练语料进行处理：

? 你帮助他们打扫房子吗?

? Do you help them to clean the room ?

这样在翻译“? 你有什么专门要参观的吗?” 这个句子的时候，句首的“?” 会在语言模型的时候更准确的约束我们的选择，事实上，采取我们的方法之后，这个句子被翻译为“? Do you have anything special to visit ?”。因为训练语料里“?” 与“Do” 一起出现的次数更多，所以它们在一起的语言模型得分会比“?” 与“there” 高。这样层次短语模型就会做出更正确的选择。另一方面在抽规则的时候也会抽出更多利用到语气信息的规则，比如“? 你--->? Do you”。最后我们用实验证明了这个方法的有效性，在汉英口语翻译的大规模实验中，我们的方法在测试集上有了2.49 个BLEU 值的显著提升。

本文第二节介绍了语言模型，第三节描述了我们的方法，第四节是实验和分析，第五节是总结与未来工作。

2 背景介绍

2．1 语言模型

语言模型(language model, LM) 在自然语言处理领域占有重要的地位，尤其是在基于统计模型的语音识别，机器翻译等相关研究中的到了广泛的应用。在过去许多年里，很多学者在这方面做了大量的研究[chen,*et;al:* ,1998,1996 ; Goodman,2001; Rosenfeld, 2000; Bengio,2003;Xu *etal:*,2002;Chelba *etal :* ,2000]。

一个语言模型通常用来构建一个字符串s 的概率分布$p(s)$,$p(s)$ 衡量了一个字符串在语料中出现的概率。首先定义一个语言中所有的单词的集合V ，字符串序列s 表示为$s = x_1x_2 \ldots x_l$。对于一个变量系列$X_1;X_2; \ldots X_l$，每一个变量都是集合V 中的元素，语言模型的可以算出任何一个句子出现的概率:

$$p(X_1 = x_1, X_2 = x_2 \cdots X_l = x_l)$$

用$p(x_1x_2 \ldots x_l)$ 表示$p(X_1 = x_1;X_2 = x_2 \ldots X_l = x_l)$, 根据概率分布的链式法则就可以的到：

$$
\begin{aligned}
p(s) &= p(x_1, x_2 \cdots x_l) \\
&= p(x_1)p(x_2|x_1) \cdots p(x_l|x_1 \cdots x_l) \\
&= \prod_{i=1}^{n} p(x_i|x_1 \cdots x_{i-1}) \qquad (1)
\end{aligned}
$$

在公式1中可以看出产生第i 个词的概率是由前i- 1 个词决定的，但是随着i 的增加，很难在训练集合里面找到前i - 1 个词语的出现，目前采用的方法是认为一个词的出现只与它前面n 个词语有关，满足这种条件的称为n 元语言模型(n-gram)，通常情况下n 取值不能太大，否则就会出现自由参数太多的情况。对于n 元语法模型，一个句子s 出现的概率就为：

$$
p(s) = \prod_{i=1}^{l} p(x_i|x_{i-1} \cdots x_{i-n}) \qquad (2)
$$

公式2说明了在n-gram 认为，句子中一个词的出现概率只与它前面的n 个词语有关系。

3 模型简介

层次短语模型可以自动的从双语句对中抽取形式语法，而不需要任何语言学的假设和数据标记，所以使用方便，目前被广泛的应用在统计机器翻译领域。由于抽取的能够捕捉短语之间的调序，所以对于长距离调序有一定的处理能力，但是由于语言的复杂性，这种形式化文法在处理长距离调序的时候还是存在一些问题。以汉英翻译为例，这里具体分析这样的一个句子:

他喜欢吃苹果吗?

对于这个例子有下面的几个可能用到的规则：

$$
\begin{aligned}
r_1 &: X \rightarrow < \text{他,he} > \\
r_2 &: X \rightarrow < X_1 \text{ 喜欢 } X_2, X_1 \text{ like } X_2 > \\
r_3 &: X \rightarrow < \text{吃苹果,eating apples} > \\
r_4 &: X \rightarrow < X_1 \text{ 吗 ?,Does } X_1 \text{ ?} > \\
r_5 &: X \rightarrow < \text{吗?, ?} > \\
r_6 &: X \rightarrow <?,?>
\end{aligned}
$$

对于层次短语，可能还需要用到粘和规则r_7; r_8：

$$
\begin{aligned}
r_7 &: S \rightarrow < S_1X_2, S_1X_2 > \\
r_8 &: S \rightarrow < X, X >
\end{aligned}
$$

对于这个句子可能的两种翻译过程如图1和2所示：

$$
\begin{aligned}
step1 \quad & < S_1, S_2 > \Rightarrow < S_2X_3, S_2X_3 > \quad (r_7) \\
step2 \quad & \Rightarrow < X_4X_3, X_4X_3 > \quad (r_8) \\
step3 \quad & \Rightarrow < X_5 \text{ 喜欢 } X_6X_3, X_5 \text{ like } X_6X_3 > \quad (r_2) \\
step4 \quad & \Rightarrow < \text{他 喜欢 } X_6X_3, \text{he like } X_6X_3 > \quad (r_1) \\
step5 \quad & \Rightarrow < \text{他 喜欢 吃 苹果 } X_3, \text{he like eating apples } X_3 > \quad (r_3) \\
step6 \quad & \Rightarrow < \text{他 喜欢 吃 苹果 吗 ?, he like eating apples ? > } \quad (r_5)
\end{aligned}
$$

图 1: 翻译过程 1

$$
\begin{aligned}
step1 \quad & < S_1, S_2 > \Rightarrow < X_3, X_3 > \quad (r_7) \\
step2 \quad & \Rightarrow < X_4\text{吗 ?, Does}X_4\text{? > } \quad (r_4) \\
step3 \quad & \Rightarrow < X_5\text{喜欢}X_6\text{吗 ?, Does}X_5\text{like}X_6\text{? > } \quad (r_2) \\
step4 \quad & \Rightarrow < \text{他 喜欢}X_6\text{吗 ?, Does he like}X_6\text{? > } \quad (r_1) \\
step5 \quad & \Rightarrow < \text{他 喜欢 吃 苹果 吗 ?, Does he like eating apples ? > } \quad (r_3)
\end{aligned}
$$

图 2: 翻译过程 2

可以看出图1的翻译结果为“he like eating apples ?”。这种类型的翻译结果在我们的口语翻译结果中很常见，很多疑问句的翻译结果都存在这样的问题。图2给出了另外一种正确的翻译过程，其翻译结果为“Does he like eating apples?”。

接下来我们分析这两种翻译过程的异同以及翻译模型在选择过程中可能出现的错误。通常层次短语模型在选择规则的时候主要考虑两个约束，一个是规则的翻译概率[5]，另外一个是语言模型[6]。我们从这两个方面来分析这两个结果：

1. 图1用到的非粘和规则为$\{r_1, r_2, r_3, r_5\}$，而图2用到的非粘和规则为$\{r_1, r_2, r_3, r_4\}$。可以看出，因为翻译结果非常的相似，它们对应的翻译过程也非常的相近，只有r_5和r_4这一个不同。r_4是训练语料里面“吗”对齐为“does”的时候抽到的规则，训练语料的长句子中，“吗”很可能空，这个时候就会抽出r_5这样的规则。这两个规则都具有比较高的出现概率，所以层次短语模型很难仅仅根据这少量的规则不同而选出正确的翻译。
2. 分析这两个句子的语言模型“he like eating apples ?”和“Does he like eating apples ?”，可以看出这两个句子都很通顺，比较合理。但是第一个句子存在一定的问题，就是句尾的标点符号句子的表达不对应。然而通常情况下，语言模型很难捕捉到这样的错误，因为章2.1对n-gram的描述来看，一个单词的出现只与它前面的n个词语有关系。“he like eating apples”是非常常见的句子，所以具有高的语言模型的分，后面加上“?”，由于训练语料里面，也只统计了“?”前面n个词语联合出现的概率，很可能没有与“Does”这样的疑问词没有关系，所以语言模型也很难区分出这两个句子的好坏。

综合上面的两点，可以看出，一般情况下，层次短语翻译模型在处理口语疑问句的时候，对正确和错误的翻译区分度不是很好，造成翻译结果有可能出错。为了提高翻译的准确率，本文考虑可以增强句子的语气信息，把句尾的标点符号提前到句首，这样翻译的时候就会很好的利用到这个信息。还是以图1和图2为例，在句子首部增加了“?”之后，要翻译的句子就成了“？喜欢吃苹果吗?”，利用规则r_6，翻译句首的“?”，其它

的部分翻译过程不变。这样就有了两个翻译结果“? Does he like eating apples ?” 和“? he like eating apples ?”，从语言模型的角度来看，在训练语料中“?”后面跟“Does”的概率显然比“?” 后面跟“he” 的概率大很多，其它部分语言模型得分都差不多，那么显然用了我们的方法之后调序正确的句子会的分会增加，被选中的概率也大大增大了，所以翻译系统的质量有了明显的提升。事实上，在训练语料的句首增加了“?” 之后，也会抽出更多有效的规则，比如“?--->? Does”，这样弥补了在长句中“Dose” 可能对空的情况，更有利于系统选择正确的翻译结果。但是对系统翻译结果的提升起最决定性作用的还是语言模型，在接下来在实验中可以看出来。

4 实验数据、结果和分析

4.1 实验数据

我们从沪江网、人人影视字幕组挖掘了500 万左右规模的平行句对，从里面随机抽出了4000 句对进行人工翻译后，分别取2000 句作为开发测试集。使用GIZA++(Och and Ney,2003)生成汉英词对齐文件，用层次短语模型(Chiang and Huang,2007) 进行解码。

4.2 实验结果

为了确定影响翻译结果的主要因素，我们做了5 组对比实验, 测试集在翻译的时候如果加了标点符号，在最后算BLEU 的时候都去掉。结果如表1所示, 下面列举了五组实验的实验条件。

表格 1 不同测试方法对结果影响

实验条件	Test1	Test2	Test3	Test4	Test5
BLEU（%）	29.96	30.52	**32.45**	31.58	32.13

test1 基准对比系统, 不添加任何额外信息

test2 只把训练集、测试集、和开发集的疑问问句$_9$句末标点复制到句首，重新抽规则和训练语言模型，最后进行解码。

test3 把训练集、测试集和开发集的所有句子的标点都复制到句首，重新抽规则和训练语言模型，最后进行解码。

test4 只把第一组的语言模型替换成第三组的，作为新的一组重新进行解码。

test5 把第四组的测试集、和开发集句尾的标点复制到句首，重新进行解码。

4.3 实验分析

如表1所示，第三组实验，当所有语料都经过句末标点复制到句首的预处理之后，效果最好。第二组实验只将问句的标点提前，可能对翻译造成干扰，许多陈述句的翻译也发生了变化。第五组实验探究了影响实验效果的决定性因素是语言模型，可以看出，使用基准对比系统的规则表，但是使用增加了标点的语言模型，相对于基准对比系统也有2.17 个BLEU 值的提升，只比test3 低了0.32 个BLEU 值。

限于篇幅，表2展示了采用我们的方法之后，在测试集上翻译结果与使用之前的差别。可以看出我们的方法对疑问句的翻译有非常大的改进。对其它语气的句子也有提高，比如第4 句。疑问词的使用在语言模型的约束下也更准确了，比如第2 句。问号的加入对翻译模型引入了更多的规则，比如第11 句。对于疑问词的顺序调整也非常的明显，比如第1、6、7、8、9 句等。

5 总结与未来工作

针对层次短语模型在翻译汉英口语疑问句上存在的丢词和乱序现象，本文提出了一种将语料句末标点符号复制到句首以增加语气信息的方法，在大规模口语汉英翻译的实验中BLEU 显著提升了2.49 个点。通过分析和实验我们发现了引起翻译质量上升的主要

原因是起约束作用的语言模型引入了新的语气信息。但是，我们目前的实验仅仅局限在口语领域，对于语气信息的提取也比较粗糙，只是将句末标点简单的复制到句首。接下来，我们将细化这个工作，对句子的语气信息做更详细的分类，比如一般疑问句，特殊疑问句等，因为不同类型的疑问句事实上也具有不同的语法现象。将它们自动分类，做相应的处理，对于翻译结果会有更大的改进。

表格 2 实验结果对比

	原文	原始翻译	改进后翻译
1	去北京最好的时间是什么时候?	Beijing's best of time?	When is the best time to Beijing?
2	他们住在哪里?	Where they live?	Where do they live?
3	你的爱好是什么?	Your favorite?	What's your hobby?
4	太激动人心了!	too exciting !	How exciting !
5	你为什么这么忙?	Why do you busy?	Why are you so busy?
6	我是你的老板吗?	I am your boss?	Am I your boss?
7	我能骂人吗?	I can swear at people?	Can I swear at people?
8	你觉得我学些什么好呢?	You think I learn anything better?	Do you think I learn anything better?
9	能把窗户关起来吗?	Can up the window?	Can you shut up the window ?
10	我在何处可取得行李?	I where can get luggage?	Where I can get luggage?
11	请问饭厅在哪?	Is where the dining room?	Excuse me, where is the dining room ?

参考文献

[1] David Chiang,Huang Liang. 2005. *A hierarchical phrase-based model for statistical machine translation*. ACL, pages 263–270.

[2] Och F J, Ney H.2004a. *The Alignment Template Approach to Statistical Machine Translation*. Computational Linguistics,30(4): 417-449.

[3] Och F J, Ney H. 2003. *A System Comparision of Various Statistic Machine Translation.* Computational Liguistics,21(9),:19-51

[4] ChenKL,GoodmanJT.1998.*An Empirical Study of Smoothing Techniques for Language Model*. TechnicalReportTR-10-98.

[5] Goodman J. 2001. *A Bit of Progress in Language Modeling*. Computer Speech and Language,15(4),pages 403-434

[6] Rosenfeld R. 2000. *Two Decades of Statistical Language Modeling:Where Do We Go from Here*. Proceedings of threeIEEE, Vol.88,No.8.pages 1270-1278

[7] BengioY,DucharmeR,VincentP,JauvinC.2003.*A Neural Probabilistic Language Model*.Journal of Machine Learning Research(3):1137-1155

[8] Chelba C,jeline F. 2000. *Structured Language Modeling*. Computer Speech and Language, 14(4):283,332

[9] Xu P, Chelba C, Jelinek F. 2002. *A Study on Richer Syntactic Dependencies for Structured Language Modeling* In:Proceeding of the 40^{th} ACL,Philadelphia,July.pages191-198