日常的口语语句中，语气词、重复、冗余、省略等非规范现象常常出现，并且经过语音识别后还存在很多的识别错误，对后续的机器翻译系统会产生很大影响，主要从三方面进行口语标准化，包括修正断句错误，ASR 错误，口语不流畅和语法不规范。

标注数据的原则：

a） 改的前后的句子使用同一种语言；

b） 改写前后的句子具有相同的语义；

c） 改写后的句子应尽量比改写前的句子简化。

断句

由于语言所的这批数据，断句存在不规范的问题，所以在标数据的时候要进行人工检测并修正断句。断句的原则

1） 每一条数据都不能太长，尽量使长度不超过 50 个字。

2） 要尽量保证每一条数据的语义完整性

3） 修正过程中，遇到逗号、句号的错误使用，也要修正。

例如下面的原始数据：

恩它它能让我们更好的包容一切，而一个具有心胸狭窄的人。

在社会往往是得不到别人尊重的，恩蓝色它更是一种天空。

的色彩它能把整个世界都包括覆盖起来。

修正断句：

它能让我们更好的包容一切。

而一个心胸狭窄的人，在社会往往是得不到别人尊重的。

蓝色更是一种天空的色彩它能把整个世界都覆盖起来。

ASR 错误

语音识别错误：汉语连续语音识别结果中的错误，

按性质分，有：

（1） 替换错误

原句：中日友好合作关系

识别结果：终日友好合作关系

（2） 删除错误

原句：学习和借鉴外国的经验

识别结果：学习借鉴外国的经验

（3） 插入错误

原句：他没有丰厚的财产

识别结果：他没有丰厚财产

按原因分，有：

（1） 同音字/词

原句：我们是学生

识别结果：我们市学生

（2） 近音字/词

原句：与家人团聚的温暖幸福

识别结果：与家人团聚的温暖信服

（3） 由于说话人速度快，或环境噪声的影响，造成漏音、冗余音、前后音粘连

原句：但是他留给人们的精神遗传

识别结果：但是他留给人们精神遗传

1 我想查询以下就是路附近的医院

原句：我想查询一下知春路附近的医院

2 我想到人民大学体育馆的信息

原句：我想知道人民大学体育馆的信息

3 那个是你的

原句：哪个是你的

不流畅

不流畅检测中主要是两类不流畅：Filler，Restart。

- Filler：句子中没有实际含义的词语，一般是语气词或呼应性的词，它们对句子含义没有任何贡献，实际上在中文中这类词语可以用一个有限集合定义出来，是可枚举的。

下表列举了部分这类词：

单字词	唉,那,嗯,喂,啊,噢,恩,呀,吧,呃,嘛,呢
复词	这个,那个,就说,就是说,那么,什么的,或者什么的,(所以)说,比如说,就是说,也等于（是）

在中文里这些词的出现不一定都是 Filler，存在歧义，所以需要人工标注。

示例：

原句：就是说我们国家现在的发展很快

顺滑后：我们国家现在的发展很快

- Restart：口语中存在的重复，冗余，修正，不完整词语替换插入等 ，包括以下 6 类：
  1. 简单重复：这种类型修正词和被修正词是相同的，可能是由于讲话者结巴引起的，如下例所示：

  原句：我知道我知道那个手段更复杂

  顺滑后：我知道那个手段更复杂
  2. 局部重复：指单词中只有几个字被重复，被修正词是修正词的一部分，如下例所示：

  原句：我们那些农农民的儿子

  顺滑后：我们那些农民的儿子
  3. 单词置换：在这种类型中，修正词和被修正词之间存在交集，但又不完全相同，如下例所示：

  原句：我想当时的,那时候的场面会非常的火爆

  顺滑后：我想那时候的场面会非常的火爆
  4. 单词移除：被修正词完全被删除，后续没有为其校正的修正词。

  原句：但我自己感到当然也很非常沮丧啊

  顺滑后：但我自己感到当然也非常沮丧啊
  5. 单词插入：被修正词是修正词的子集，修正词在被修正词的基础上进行了扩充。

  原句：罗家虹他也有他自己家也有床

  顺滑后：罗家虹他自己家也有床
  6. 复杂类型：由以上五种简单类型通过嵌套组成的复杂结构

  原句：对一月一一月底回去的

  顺滑后：对一月底回去的

不规范

非规范句子结构分为以下 6 种：

1. 搭配不当：搭配不当包括主谓搭配不当，动宾搭配不当，定、状、补与中心词搭配不当和主宾意义搭配不当等多种情况。例如：

   1　这个公司的销售量大踏步地发展。（主谓搭配不当）

   2　他主动挑起了这项任务。（动宾搭配不当）

   3　坚实的基础知识为他进一步学习提供了条件。（定语与中心词搭配不当）

   搭配不当的句子一般没有句法上的错误，只有对相应的词进行语义检查才能够分辨。

原始口语	有道翻译
这个公司的销售量大踏步地发展。	The sales of this company have been greatly improved.
他主动挑起了这项任务。	He initiated the task.
坚实的基础知识为他进一步学习提供了条件。	Solid basic knowledge provided the conditions for his further study.

2. 句子残缺：指在对话过程中，说话人根据上下文语境都明白而为了话语的方便性而无需再说出的字词信息。不符合省略条件而缺少某种句子成分，常常引起句子结构不完整，表达意思不准确。还有些省略尽管在语法上符合，但由于省略成分太多，导致句子表达意思不完整。例如：

   1　由于软件技术的提高，为微机普及提供了条件。（主语短缺）

   2　学习计算机的过程是不会到会循序渐进。（宾语残缺）

   3　（在公共汽车上售票员说）西单下。

   4　爱信不信。

   5　航天桥附近那个（加油站）

   6　十三班（有没有人参加）呢

原始口语	有道翻译
由于软件技术的提高，为微机普及提供了条件。	Because of the improvement of software technology, it provides the conditions for the popularization of microcomputers.
学习计算机的过程是不会到会循序渐进。	The process of learning a computer is not gradual.
西单下。	Xidan.
爱信不信。	Believe it or not.
航天桥附近那个	The one near the space bridge.
十三班呢	Ten class three?

3. 成分冗余：句子中含有多余成分，尤其是主要成分，会引起句子意思混乱或含混不清。还有一种情况是当说话人思维不连贯而在话语中添加一些辅助词语以增加语气的节奏感，或者由于意外而导致的语句中含有不符合语法规则的词语重复现象，这些多余的字或词语造成了冗余，不包括“看看”、“谢谢”这类复词，去掉它们对整个语句的原意没有影响。例如：

   1　我们青年人我们必须刻苦学习。（主语多余）

2 今天下午少先队员为英雄举行了献花。（谓语多余）

3 我想问一下清华附近的啊麦当劳在哪

4 嗯那个香山饭店的价位怎么样

5 我想问一下中关村中关村周围规模较大的医院都有哪些

6 下礼拜下礼拜二三吧好吗

原始口语	有道翻译
我们青年人我们必须刻苦学习。	We young people must study hard.
今天下午少先队员为英雄举行了献花。	The young pioneers presented flowers for the hero this afternoon.
我想问一下清华附近的啊麦当劳在哪	I want to know where is McDonald's near tsinghua university?
嗯那个香山饭店的价位怎么样	Well, what about the price of the fragrant hill hotel?
我想问一下中关村中关村周围规模较大的医院都有哪些	I would like to ask about the larger hospitals around zhongguancun in zhongguancun.
下礼拜下礼拜二三吧好吗	How about next Tuesday, Wednesday?

4. 语序不当：如果严格地按汉语语法规定，就会发现定语与中心词、定语与状语以及多层定语之间和多层状语之间语序不当的情况非常普遍。例如：

1 前面那间屋，他们两个住。（定语与中心词位置颠倒）

调序后：他们两个住，前面那间屋。

2 它晚上要跟你床上一起睡。（定语与中心词位置颠倒）

调序后：它晚上要跟你一起睡床上。

3 或者说是这边谈不拢，那边谈不拢，在价钱问题上。

调序后：或者说是在价钱问题上这边谈不拢，那边谈不拢。

4 上个星期在的一场比赛，被弗勒姆竟然压着打，

调序后：上个星期在的一场比赛， 竟然被弗勒姆压着打，

5 黄易的小说，介于武侠和虚幻之间，也有他很独到的地方，当然。

调序后：黄易的小说，介于武侠和虚幻之间，当然也有他很独到的地方。

6 家务活基本上也是都她干的。

调序后：家务活基本上也都是她干的。

7　都是上海人自己灌注自己感情在。

调序后：都是上海人自己在灌注自己感情。

8　比如说,中国人为什么和日本有像深仇大恨一样的。

调序后：比如说,中国人为什么和日本像有深仇大恨一样的。

9　其实就是一个环境的问题，我觉得。

调序后：其实我觉得就是一个环境的问题。

10　他对我们几个班的干部说：你们二班……（两个定语次序不当）

调序后：他对我们班的几个干部说：你们二班……

原始口语	有道翻译
前面那间屋，他们两个住。	In the front room, they live in two.
它晚上要跟你床上一起睡。	It sleeps with your bed at night.
或者说是这边谈不拢，那边谈不拢，在价钱问题上。	Or we can't talk about it. We can't talk about it.
上个星期在的一场比赛，被弗勒姆竟然压着打，	In a game last week, he was crushed by flom,
家务活基本上也是都她干的。	She did all the housework.

5. 句式杂糅：句式杂糅常使句子成分残缺或句子结构混乱。例如

1　听了他的报告，对我启发很大。（两句杂糅）

2　获奖后，我们有既光荣又愉快的感觉是很难形容的。（前后牵连）

原始语料	有道翻译
听了他的报告，对我启发很大。	I was inspired by his report.
获奖后，我们有既光荣又愉快的感觉是很难形容的。	After winning the prize, we have a glorious and pleasant feeling that is hard to describe.

6. 口语词汇：指句子中使用了一般只在口语或方言中使用的词。例如：

1　这样也蛮好的

2　昨儿就到学校了

3　等你老半天儿了

原始语料	有道翻译
这样也蛮好的	That's fine.
昨儿就到学校了	I went to school yesterday.
等你老半天儿了	I'll be waiting for you.

7. 含有歧义：句子歧义包括句子中含有多义词或多义短语、词与词的承接关系不明确等多种情况。例如：

   1　飞着去还是坐着去？（意思是：坐飞机去还是乘火车去？）

   2　（朋友问）家里那位怎么样？（意思是：你爱人还好吗？）

原始语料	有道翻译
飞着去还是坐着去?	Flying or sitting?
家里那位怎么样?	How about the family?

参考文献

[1]吴双志. 语音翻译中口语文本规范化的研究[D]. 哈尔滨工业大学, 2015.

[2]宗成庆.音字转换与句子规范化处理研究[D]. 中国科学院研究生院(计算技术研究所), 1998.

[3]吴斌. 语音识别中的后处理技术研究[D]. 北京邮电大学, 2008.

[4]徐波. 基于条件随机场的口语规范化处理研究[D]. 南京理工大学, 2009.