**RFC文件的特定Test Case生成方案**

**及工作进展**

1. **研究背景**

本工作主要使用NLP语义理解技术实现从RFC文档的不确定性语句中自动生成特定Test Case。

特定Test Case生成的前置条件是有定义好的Rule文件，如图1所示为“Header\_Copy”的示例。在Rule文件中需要明确定义Test Case的类型（如“Header\_Copy”）和属性名称（如“name”、”num”、”value”、”style”）。

num值：value[]中填多少个值是不确定的。<num个值就用默认值补到num个；=num个值最好；>num个值就从中取num个值的排列



图1. 一条Rule的示例

例如给定图2所示句子，我们使用自然语言处理及语义理解技术，达到生成如图3所示的Test Case结果的目标。 从文本段落中。。。。。。。。。。。。。



图2. RFC中的示例语句

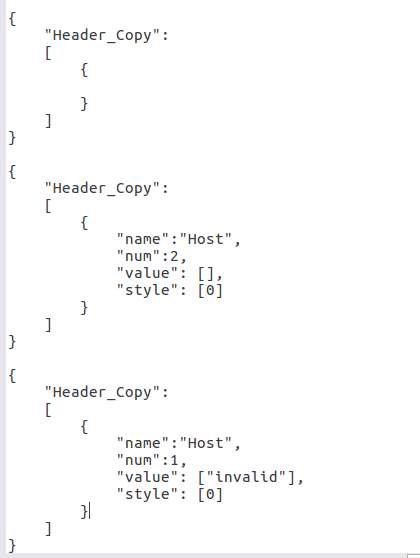


图3. Test Case示例结果

1. **Test Case生成方案**

从RFC中生成Test Case的整体方案流程如下：

1. **不确定性语句的筛选**

* 对整个RFC文档进行分句；
* 筛选包含大写情态动词的句子（RFC文档中情态动词均为大写）

1. **对长句的重写**

如图2所示的长句句式结构往往比较复杂，直接对复杂句进行语义理解是非常有难度的。在本工作中，我们使用了Spacy[1]作为依存句法关系分析的工具（Spacy在英文文本的处理上具有较高准确率），通过分析并列依存关系，独创性地将复杂句重写为若干简单句。

* 对句子进行句法依存分析；
* 判断句子是否存在并列成分，如果存在，则继续执行下面的处理步骤，否则退出当前环节；
* 通过分析句子成分之间的并列依存关系，设计算法实现并列元素的拆分和语句的重写（图2 中的句子重写后如图4所示）。

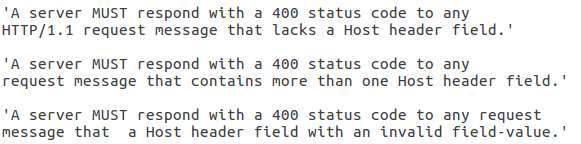


图4. 长句重写后的句子示例

1. **Rule Type映射**

Rule Type映射是Test Case生成的第一步，只有在确定Rule Type后，我们才能进行特定范围内的属性类别和属性值的设置。

我们定义的Rule Type可能有多个，一个简单句可能与其中一种类别匹配，也可能不与任何类型匹配，因此这不是一个完全的分类任务。我们采取特征词映射来确定句子所属于的Rule Type。具体实现方法描述如下：

* 人工设定种子词。例如“header\_copy”的种子词为multiple, more than one, lack等。
* 种子词扩展。人工设定的种子词数量有限，不能将相近语义的词汇列举完。我们通过预训练的词向量模型扩展出与种子词相似的词语，为确保扩展质量，可以再进行人工筛选。
* 为每个Rule Type构造特征词表（由种子词及扩展词构成）。
* 计算待处理语句与Rule Type特征词表的匹配得分，通过设定得分的阈值来判定属于哪一个Rule Type或者不属于任何Rule Type。

1. **属性值映射**

每一个Rule Type都有特定的若干个属性，我们需要理解待处理句子，从中分析并得到每个属性对应的属性值。同Rule Type映射相似，我们经过种子词设定、种子词扩展为每一个属性构造一个特征词映射字典。通过特征词映射字典中关键词的匹配可以找到句子中出现的匹配词语。

**三、 工作进展情况**

1. “不确定性语句筛选”的模块已经完成

* 对RFC2616.txt和RFC7230.txt完成了文本分句，并分别筛选了720条和207条带有情态动词的不确定性语句。

1. “对长句重写”的模块正在进行中

* 已经完成了使用spacy对语句的句法分析
* 已经完成了算法设计
* 正在实现设计的算法

注：“长句重写”可以为Rule Type映射和属性映射提供方便，是一项必须的工作。Spacy是英文NLP中最好的工具，官方报告在句法依存中的准确率为94.48%。虽然有如此高的准确率，但仍有误差存在。因此由句法分析结果的误差导致的长句重写错误也是在实际中会出现的情况。

几个属于header\_copy的句子：

It MUST be possible to combine the multiple header fields into one "field-name: field-value" pair, without changing the semantics of the message, by appending each subsequent field-value to the first, each separated by a comma.

A recipient MAY replace any linear white space with a single SP before interpreting the field value or forwarding the message downstream.

注释： LWS （linear white space）