STA5007 Advanced Natural Language Processing

Homework 1 Report

12112627 李乐平

Question 1

话题 1: Text Simplification 文本简化

简介: 文本简化是在维持文本原意的条件下，减少其词汇及句式结构复杂性使之更容易理解和阅读的任务。文本简化有相当多而重要的社会应用，例如帮助失语者、诵读困难者、自闭症患者等具有认知障碍的人或非母语者和儿童等具有阅读困难的人。

输入输出: 模型的输入为文本，输出为简化后的文本。

评估指标: 评价文本简化效果的主要指标是文本可读性。文本可读性的具体评估指标有很多，常用的评估指标有 FKGL（Flesch-Kincaid Grade Level）、BLEU（Bilingual Evaluation Understudy）、SARI（System output Against References and against the Input sentence）及其改进 EASSE（Easier Autometic Sentence Simplification Evaluation）等。

下面简要介绍这几种指标：

· FKGL

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

FKGL 是一种简单的衡量公式，其被广泛应用于文本可读性的对比。但也有学者认为这种方法应该被淘汰（https://aclanthology.org/2021.gem-1.1.pdf），理由是在其他几种指标没有较大改变的情况下，此分数能通过某些操作轻易地大幅提高。

· BLEU

BLEU 的计算方式比较复杂，其主要思想是比较连续比较机器生成的文本（Candidate）与参考的人工文本（Reference）的 n 元词序列之间的相似性，并引入过短惩罚系数（Brevity Penalty）以平衡文本长度差异。

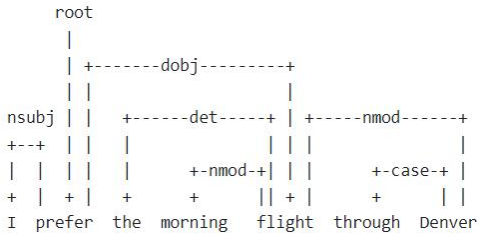
· SARI

在 BLEU 的基础上，SARI 还额外考虑了与原始文本的对比。其主要思想是奖励有益的增加和保留操作，并惩罚过度的删除操作。

不妨选择在 TurkCorpus 数据集上的 SARI(EASSE>=0.2.1)指标，表现最好的两个模型是 MUSS(BART+ACCESS supervised)和 Control Prefixes(BART)，它们分别获得了 42.53 和 42.32 的评分。

话题 2: Dependency Parsing 依存分析

简介：依存分析旨在提取句子中词与词之间的依存关系以表现其语法结构。例如:



句子「I prefer the morning flight through Denver」被提取为了一棵有向树。曲线从中心词（Head，用「|」连出）指向依存词（Dependency，用「+」连入）。

输入输出: 模型的输入为一句话（即一串有序的字符），输出为表示各词关系的依存树。

评估指标: 常用的评估指标有 LAS(有标签标注准确率，Labeled Attachment Score)、UAS(无标签标注准确率，Unlabeled Attachment Score)等。LAS 关注有多少单词被正确地连到了其对应的中心词，同时还被打上了正确的标签；而 UAS 则不关注其标签是否正确。

以在经典数据集 Penn Treebank 上的无标签标注准确率为指标，表现最好的模型是 Label Attention Layer + HPSG + XLNet，准确率 97.42；次好的模型为 DMPar + XLNet，准确率 97.30。

话题 3: 图像描述 Image Captioning

简介: 图像描述旨在将输入的图片转化为描述其内容的文字。

输入输出: 模型的输入为图片，输出为描述图片内容的字符串。

评估指标: 评估指标有很多种，常用的有 BLEU、CIDEr(Consensus-based Image Description Evaluation)、SPICE 等。其中，CIDEr 的主要思想是对于参考文本中出现频率低的关键词赋予高权重，以评价输出是否把握住了关键信息；SPICE 则更多考虑了参考文本和生成文本中的依存关系，一定程度上规避了少量关键信息被替换或缺失时在其他指标上的体现不够明显的缺陷。

在经典数据集 COCO Captions 下以 CIDEr 为指标，表现最好的模型是 mPLUG，得分 155.1；次好的模型为 OFA，得分 154.9。

Question 2

代码见「12112627_李乐平_ass1.ipynb」；

运行结果见「12112627_李乐平_ass1 - Jupyter Notebook.pdf」。

参考了 ChatGPT 3.5 的回答。Prompt 如下：

How to train a sentencepiece model with vocabulary size = 3000?

How to train word embedding using skip-gram method with FastText toolkit?

How can the training connected with the sentencepiece preprocessing mentioned above?

训练出来的模型较大，因此不上传在 BB 上，通过运行代码可以直接获得。